UNIVERSITY OF CALIFORNIA

Los Angeles

Learning in Large–Scale Games and Cooperative Control

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Mechanical Engineering

by

Jason Robert Marden

2007

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **2007**	2. REPORT TYPE	3. DATES COVERED **00-00-2007 to 00-00-2007**

4. TITLE AND SUBTITLE **Learning in Large-Scale Games and Cooperative Control**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California, Los Angeles,Department of Mechanical Engineering,Los Angeles,CA,90095**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S) 11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Many engineering systems can be characterized as a large scale collection of interacting subsystems each having access to local information, making local decisions having local interactions with neighbors, and seeking to optimize local objectives that may well be in conflict with other subsystems. The analysis and design of such control systems falls under the broader framework of ?complex and distributed systems?. Other names include ?multi-agent control,? ?cooperative control,? ?networked control,? as well as ?team theory? or ?swarming.? Regardless of the nomenclature, the central challenge remains the same. That is to derive desirable collective behaviors through the design of individual agent control algorithms. The potential benefits of distributed decision architectures include the opportunity for real-time adaptation (or self-organization) and robustness to dynamic uncertainties such as individual component failures, non-stationary environments, and adversarial elements. These benefits come with significant challenges, such as the complexity associated with a potentially large number of interacting agents and the analytical difficulties of dealing with overlapping and partial information. This dissertation focuses on dealing with the distributed nature of decision making and information processing through a non-cooperative game-theoretic formulation. The interactions of a distributed/multi-agent control system are modeled as a noncooperative game among agents with the desired collective behavior being expressed as a Nash equilibrium. In large scale multi-agent systems, agents are inherently limited in both their observational and computational capabilities. Therefore, this dissertation focuses on learning algorithms that can accommodate these limitations while still guaranteeing convergence to a Nash equilibrium. Furthermore, in this dissertation we illustrate a connection between the fields of game theory and cooperative control and develop several suitable learning algorithms for a wide variety of cooperative control problems. This connection establishes a framework for designing and analyzing multi-agent systems. We demonstrate the potential benefits of this framework on several cooperative control problems including dynamic sensor coverage, consensus, and distributing routing over a network, as well as the mathematical puzzle Sudoku.

15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT **Same as Report (SAR)**	18. NUMBER OF PAGES **184**	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**			

The dissertation of Jason Robert Marden is approved.

Gürdal Arslan

Robert M'Closkey

Jason L. Speyer

Jeff S. Shamma, Committee Chair

University of California, Los Angeles

2007

To my family . . .

whose love and support have guided me through the years

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Words cannot even begin to express the gratitude that I feel towards my family, friends, colleagues and mentors for their unconditional support, encouragement and guidance over the past six years. This thesis could not have been completed without all of you.

When I started graduate school at UCLA, I knew very little about my new advisor, Jeff Shamma. All that I knew about Jeff was that I enjoyed talking to him and that he promised me a free round of golf, which I am still waiting for to this day. Six years later, I now realize how extremely fortunate I am to have had the opportunity to get to work with Jeff. Jeff is one of the most sincere and caring individuals that I have ever met in my life. Jeff took me under his wing like a father would do to his own son. He guided, mentored, and provided me with a wealth of opportunities that ultimately changed the direction of my life forever. I can never repay Jeff for all that he has done for me. Thank you, Jeff. I am so proud to call you my academic father.

I was extremely fortunate to not have just one advisor, but rather two. My second advisor, Gürdal Arslan, or “G” as the local Hawaiians call him, came into my life about four years ago at a time when I began to question both life and my ambition for research. Gürdal always made time for me, challenged me, and ultimately helped me grow as both an individual and a researcher. Gürdal, through his “harnessed optimism” and attention to detail, i.e., “first sentence, third paragraph... third sentence, third paragraph...,” taught me discipline, patience, and precision in the research process. Gürdal is now one of my best friends and will continue to be for the rest of my life. Thank you for everything G. Andy’s on me!

During my graduate school experience, I became interested in the field of learning in games largely due to the work of H. Peyton Young. While attending a game theory conference in New York in July of 2006, Peyton agreed to meet with me to discuss

my research. We went on to have several interesting research conversations in addition to writing a journal paper together. Thank you Peyton for making the time to work with and mentor a non-economic graduate student. I have thoroughly enjoyed our conversations and I hope that we can continue to work together for many more years to come.

I have had the pleasure of learning from and interacting with several fantastic professors over the past six years, including but not limited to Munther Dahleh, Robert M'Closkey, Jason Speyer, T.C. Tsao, and Web Marner. In addition to academia, I have also had two fantastic mentors while working in industry, Greg Becker and John Abedor. I want to thank all of you for your time and support. I would also like to acknowledge the funding sources that made my research experience possible: NSF grants #CMS–0339228 and #ECS-0501394, ARO Grant #W911NF–04–1–0316 and AFOSR grant #FA9550-05-1-0239.

Sharing the graduate school experience with my lab mates and fellow colleagues has given me a lifetime of fantastic memories. Our research conversations over boba and coffee breaks have contributed significantly to this dissertation in addition to my personal growth. To my main man George, or the "Algebra Guy" as the women call him, your dedication and work ethic have inspired me tremendously over the past five years. Ibrahim, you have definitely taught me valuable lessons in both life and the art of debating. Whenever I hear the words, "well not exactly ...," I will always think of you. Shalom, you've always been there to offer me your "sound" advice, which thankfully I have not followed. Talk about .... in a bag. Mike, you were always the one person that I could always talk to about my graduate school experiences. I really felt that we were on the same page with so many issues. To Li Na and Clay, as you start your doctoral journey, you are a reminder that the academic life cycle continues and the future is so bright. And lastly to Dr. Jonathan Wolfe, whose life ended all too soon, your love for life and your optimism will be an inspiration to me for the rest of my life. You will always be remembered.

My friends and family were the reason why I was able to persevere through the six year grind of graduate school. I am so fortunate to be blessed with such a wonderful group of friends who have shared this experience with me over the years. To Nick, my brother Brian, Big Boy, my cousins Jon and Michael, Annie, Tommy, Mark, Steve, Jeff, Dustin, and Terri, thank you for always being there for me. Your support, encouragement, and perpetual jokes about my lifelong student status will always be remembered. I have thoroughly enjoyed sharing this experience with each one of you and I am so grateful to have all of you in my life.

My family has been my pillar of strength for my entire life, especially during my graduate studies. To Mom and Dad, you have given me so much over the years that I do not even know where to begin. You have always been there for me, cared for me and loved me. Both of you are the reason why I am here today. This dissertation is truly a tribute to your dedication in raising me. You have inspired me to be a better son, husband, and hopefully one day, a father. I love you with all my heart and I am so proud to have you both as my parents. To my Grams, your warm heart and kindness will live with me for the rest of my life. You are a constant reminder of the true essence of life – love and happiness.

To my beloved wife Nancy, you are the love of my life. Thank you for standing with me for the past five years as I pursued my dreams. Your unconditional love and support has provided me with the strength to overcome each obstacle. You have inspired me in so many ways by your kindness, generosity, compassion and sincerity. You have opened my eyes and my heart to a more fulfilling and meaningful life. You have truly been my guiding light. I have grown so much as person with you by my side and I am so blessed to have you in my life.

Lastly, I want to thank God for making all of this possible.

PUBLICATIONS

J. R. Marden and G. Arslan and J. S. Shamma. “Joint Strategy Fictitious Play with Inertia for Potential Games.” In *Proceedings of the 44th IEEE Conference on Decision and Control*, pp. 6692–6697, December 2005.

J. R. Marden and G. Arslan and J. S. Shamma. “Regret based dynamics: Convergence in Weakly Acyclic Games.” In *Proceedings of the 2007 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May 2007.

J. R. Marden and G. Arslan and J. S. Shamma. “Connections Between Cooperative Control and Potential Games Illustrated on the Consensus Problem.” In *Proceedings of the 2007 European Control Conference (ECC ’07)*, July 2007.

G. Arslan, J. R. Marden, and J. S. Shamma. “Autonomous vehicle-target assignment: a game theoretical formulation.” *ASME Journal of Dynamic Systems, Measurement and Control*, 2007.

ABSTRACT OF THE DISSERTATION

Learning in Large–Scale Games and Cooperative Control

by

Jason Robert Marden

Doctor of Philosophy in Mechanical Engineering

University of California, Los Angeles, 2007

Professor Jeff S. Shamma, Chair

Many engineering systems can be characterized as a large scale collection of interacting subsystems each having access to local information, making local decisions, having local interactions with neighbors, and seeking to optimize local objectives that may well be in conflict with other subsystems. The analysis and design of such control systems falls under the broader framework of “complex and distributed systems”. Other names include “multi-agent control,” “cooperative control,” “networked control,” as well as “team theory” or “swarming.” Regardless of the nomenclature, the central challenge remains the same. That is to derive desirable collective behaviors through the design of individual agent control algorithms. The potential benefits of distributed decision architectures include the opportunity for real-time adaptation (or self-organization) and robustness to dynamic uncertainties such as individual component failures, non-stationary environments, and adversarial elements. These benefits come with significant challenges, such as the complexity associated with a potentially large number of interacting agents and the analytical difficulties of dealing with overlapping and partial information.

This dissertation focuses on dealing with the distributed nature of decision making and information processing through a non-cooperative game-theoretic formulation. The interactions of a distributed/multi-agent control system are modeled as a non-cooperative game among agents with the desired collective behavior being expressed as a Nash equilibrium. In large scale multi-agent systems, agents are inherently limited in both their observational and computational capabilities. Therefore, this dissertation focuses on learning algorithms that can accommodate these limitations while still guaranteeing convergence to a Nash equilibrium. Furthermore, in this dissertation we illustrate a connection between the fields of game theory and cooperative control and develop several suitable learning algorithms for a wide variety of cooperative control problems. This connection establishes a framework for designing and analyzing multi-agent systems. We demonstrate the potential benefits of this framework on several cooperative control problems including dynamic sensor coverage, consensus, and distributing routing over a network, as well as the mathematical puzzle Sudoku.

CHAPTER 1

Overview

Many engineering systems can be characterized as a large scale collection of interacting subsystems each having access to local information, making local decisions, having local interactions with neighbors, and seeking to optimize local objectives that may well be in conflict with other subsystems. A representative sampling includes autonomous vehicle teams, cooperative robotics, distributed computing, electronic commerce, wireless networks, sensor networks, traffic control, social networks, and combat systems.

The analysis and design of such control systems falls under the broader framework of “complex and distributed systems”. Other names include “multi-agent control,” “cooperative control,” “networked control,” as well as “team theory” or “swarming.” Regardless of the nomenclature, the central challenge remains the same. That is to derive desirable collective behaviors through the design of individual agent control algorithms. The potential benefits of distributed decision architectures include the opportunity for real-time adaptation (or self-organization) and robustness to dynamic uncertainties such as individual component failures, non-stationary environments, and adversarial elements. These benefits come with significant challenges, such as the complexity associated with a potentially large number of interacting agents and the analytical difficulties of dealing with overlapping and partial information.

This dissertation focuses on dealing with the distributed nature of decision making and information processing through a non-cooperative game-theoretic formulation.

The interactions of a distributed/multi-agent control system are modeled as a non-cooperative game among agents, with the desired collective behavior being expressed as a Nash equilibrium. The emphasis is on simple learning algorithms that guarantee convergence to a Nash equilibrium. Furthermore, the algorithms must have minimal computational requirements to accommodate implementation in a wide variety of engineered systems.

The need for simple learning algorithms can be motivated by looking at the problem of distributed routing over a network. In such a problem, there is a large number of self interested players seeking to utilize a common network. Since the available resources in the network are finite, players' objectives are very much in conflict with one another. The sheer quantity of available information makes centralized dissemination or processing infeasible. When modeling the players' interaction as a non-cooperative game, the central issue involves how players make decisions. Or more precisely, what information do players need to base their decisions on so as to guarantee some form of a collective behavior? For example, does each player need to know the routing strategies of all other players or would some form of aggregate information be acceptable?

Motivated by the inherent information restrictions in the problem of distributed routing over networks, in Chapter 3 we consider multi-player repeated games involving a large number of players with large strategy spaces and enmeshed utility structures. In these "large-scale" games, players are inherently faced with limitations in both their observational and computational capabilities. Accordingly, players in large-scale games need to make their decisions using algorithms that accommodate limitations in information gathering and processing. This disqualifies some of the well known decision making models such as "Fictitious Play" (FP) [MS96a], in which each player must monitor the individual actions of every other player and must optimize over a high dimensional probability space.

In this chapter, we analyze the properties of the learning algorithm Joint Strategy Fictitious Play (JSFP), a close variant of FP. We demonstrate that JSFP alleviates both the informational and computational burden of FP. Furthermore, we introduce JSFP with inertia, i.e., a probabilistic reluctance to change strategies, and establish the convergence to a pure Nash equilibrium in all generalized ordinal potential games in both cases of averaged or exponentially discounted historical data. We illustrate JSFP with inertia on the specific class of congestion games, a subset of generalized ordinal potential games. In particular, we illustrate the main results on a distributed traffic routing problem.

In Chapter 4, we extend the results of JSFP by introducing an entire class of learning algorithms that can accommodate such observational and processing restrictions. To that end, we build upon the idea of no-regret algorithms [HM00] to strengthen the performance guarantees for implementation in multi-agent systems. No-regret algorithms have been proposed to control a wide variety of multi-agent systems. The appeal of no-regret algorithms is that they are easily implementable in large scale multi-agent systems because players make decisions using only regret based information. Furthermore, there are existing results proving that the collective behavior will asymptotically converge to a set of points of "no-regret" in any game. We illustrate, through a simple example, that no-regret points need not reflect desirable operating conditions for a multi-agent system.

Multi-agent systems often exhibit an additional structure, i.e., being weakly acyclic, that has not been exploited in the context of no-regret algorithms. In this chapter, we introduce a modification of the traditional no-regret algorithms by (i) exponentially discounting the memory and (ii) bringing in a notion of inertia in players' decision process. We show how these modifications can lead to an entire class of regret based algorithms that provide *almost sure* convergence to a pure Nash equilibrium in any weakly acyclic game.

The last, and most informationally restrictive, class of learning algorithms that we will consider in this dissertation are payoff based algorithms. In such a scenario, players *only* have access to (i) the action they played and (ii) the utility (possibly noisy) they received. In a transportation network, this translates to drivers only having information about the congestion actually experienced. Drivers are now unaware of the traffic conditions on any alternative routes, which was previously a requirement for the implementation of either JSFP or any regret based learning algorithm.

In Chapter 5, we focus on payoff based learning algorithms on the specific class of weakly acyclic games. We introduce three different payoff based processes for increasingly general scenarios and prove that after a sufficiently large number of stages, player actions constitute a Nash equilibrium at any stage with arbitrarily high probability. The first learning algorithm, called *Safe Experimentation*, guarantees convergence to an optimal Nash equilibrium in any identical interest game. Such an equilibrium is called optimal because it maximizes the payoff to all players. The second learning algorithm, called *Simple Experimentation*, guarantees convergence to a Nash equilibrium in any weakly acyclic game. The third learning algorithm, called *Sample Experimentation*, guarantees convergence to a Nash equilibrium in any weakly acyclic game even in the presence of noisy utility functions.

The second topic of Chapter 5 is centered around the inefficiency of Nash equilibria in routing problems. It is well known that a Nash equilibrium may not represent a desirable operating point in a routing problem as it typically does not minimize the total congestion on the network. Motivated by this inefficiency concern, we derive an approach for modifying player utility functions through tolls and incentives in congestion games, a special class of weakly acyclic games, to guarantee that a centralized objective can be realized as a Nash equilibrium. We illustrate this equilibrium refinement method on a well studied distributed routing problem known as Braess' Paradox.

In the following chapter, the focus shifts from the development of suitable learning algorithms to understanding how one would design a multi-agent systems for a cooperative control problem. In particular, how would a global planner design each agent's local utility function such that a central objective could be realized as the outcome of a repeated non-cooperative game? We seek to answer this question by highlighting a connection between cooperative control problems and potential games. This connection to potential games provides a structural framework with which to study cooperative control problems and suggests an approach for utility design. However, we would like to note that utility design for multi-agent systems is still very much an open issue.

In Chapter 6, we present a view of cooperative control using the language of learning in games. We review the game theoretic concepts of potential games and weakly acyclic games and demonstrate how several cooperative control problems such as consensus, dynamic sensor coverage, and even the mathematical puzzle Sudoku can be formulated in these settings. Motivated by this connection, we build upon game theoretic concepts to better accommodate a broader class of cooperative control problems. In particular, we introduce two extensions of the learning algorithm Spatial Adaptive Play. The first extension called *binary Restricted Spatial Adaptive Play* accommodates restricted action sets caused by limitations in agent capabilities. The second extension called *Spatial Adaptive Play with Group Based Decisions* accommodates group based collaborations in the decision making process. Furthermore, we also introduce a new class of games, called *sometimes weakly acyclic games*, for time-varying utility functions and action sets, and provide distributed algorithms for convergence to an equilibrium.

Lastly, we illustrate the potential benefits of this connection on several cooperative control problems. For the consensus problem, we demonstrate that consensus can be reached even in an environment with non-convex obstructions. For the functional consensus problem, we demonstrate an approach that will allow agents to reach consensus on a specific consensus point which is some function of the initial conditions. For the dynamic sensor coverage problem, we demonstrate how autonomous sensors can distribute themselves using only local information in such a way as to maximize the probability of detecting a particular event over a given mission space. Lastly, we demonstrate how the popular mathematical game of Sudoku can be modeled as a noncooperative game and solved using the learning algorithms discussed in this dissertation.

1.1 Main Contributions of this Dissertation

To summarize, we will now restate the main contributions of this dissertation.

- We introduce the learning algorithm Joint Strategy Fictitious Play with inertia and establish almost sure convergence to a pure Nash equilibrium in all generalized ordinal potential games in both cases of averaged or exponentially discounted historical data.

- We introduce a modification of the traditional no-regret algorithms by (i) exponentially discounting the memory and (ii) bringing in a notion of inertia in players' decision process. We show how these modifications can lead to an entire class of regret based algorithms that provide almost sure convergence to a pure Nash equilibrium in any weakly acyclic game.

- We introduce the payoff based algorithm Safe Experimentation and establish almost sure convergence to an optimal Nash equilibrium in any identical interest game.

- We introduce the payoff based algorithm Simple Experimentation and establish almost sure convergence to a pure Nash equilibrium in any weakly acyclic game.

- We introduce the payoff based algorithm Sample Experimentation and establish almost sure convergence to a pure Nash equilibrium in any weakly acyclic game even in the presence of noisy utility functions.

- We derive an approach for modifying player utility functions through tolls and incentives in congestion games to guarantee that a centralized objective can be realized as a Nash equilibrium.

- We establish a connection between potential games and cooperative control and demonstrate the potential benefits of this connection on several cooperative control problems including dynamic sensor coverage, consensus, and distributing routing over a network, as well as the mathematical puzzle Sudoku.

- We derive an equivalent definition for weakly acyclic games that explicitly highlights the connection between weakly acyclic and potential games.

- We introduce an extension of the learning algorithm Spatial Adaptive Play, called binary Restricted Spatial Adaptive Play, to accommodate restricted action sets caused by agent limitations. We establish probabilistic convergence to an action profile that maximizes the potential function in any potential game.

- We introduce an extension of the learning algorithm Spatial Adaptive Play, called Spatial Adaptive Play with Group Based Decisions, to accommodate group based collaborations in the decision making process. We establish probabilistic convergence to an action profile that maximizes the potential function in any potential game.

- We introduce a new class of games, called sometimes weakly acyclic games, for time-varying utility functions and action sets, and provide distributed algorithms for almost sure convergence to a universal Nash equilibrium.

CHAPTER 2

Background

In this section, we will present a background of the game theoretic concepts used in this dissertation. We refer the readers to [FT91, You98, You05] for a more comprehensive review.

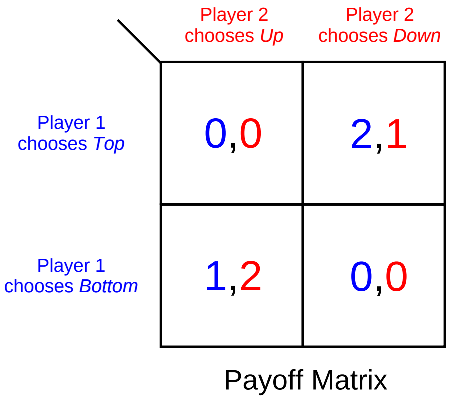
2.1 Finite Strategic-Form Games

We consider a finite strategic-form game with n-player set $\mathcal{P} := \{\mathcal{P}_1, ..., \mathcal{P}_n\}$ where each player $\mathcal{P}_i \in \mathcal{P}$ has an action set $\mathcal{A}_i$ and a utility function $U_i : \mathcal{A} \to \mathbb{R}$ where $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. We will refer to a finite strategic-form game as just a game and we will sometimes use a single symbol, e.g., G, to represent the entire game, i.e., the player set, $\mathcal{P}$, action sets, $\mathcal{A}_i$, and utility functions U_i.

An example of a two player game is illustrated in matrix form in Figure 2.1. In this game, each player has two actions or strategies and a utility function represented by the payoff matrix. Once each player has selected his action, both players receive their associated reward. For example, if player 1 choose *Top* and player 2 choose *Down*, player 1 would receive a reward of 2 while player 2 would receive a reward of 1.

For an action profile $a = (a_1, a_2, ..., a_n) \in \mathcal{A}$, let a_{-i} denote the profile of player actions *other than* player $\mathcal{P}_i$, i.e.,

$$a_{-i} = \{a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n\}.$$



	Player 2 chooses *Up*	Player 2 chooses *Down*
Player 1 chooses *Top*	0,0	2,1
Player 1 chooses *Bottom*	1,2	0,0

Payoff Matrix

Figure 2.1: Example of a Finite Strategic-Form Game

With this notation, we will sometimes write a profile a of actions as (a_i, a_{-i}). Similarly, we may write $U_i(a)$ as $U_i(a_i, a_{-i})$. Furthermore, let $\mathcal{A}_{-i} = \prod_{\mathcal{P}_j \neq \mathcal{P}_i} \mathcal{A}_i$ denote the set of possible collective actions of all players other than player $\mathcal{P}_i$ and let $\mathcal{P}_{-i} = \{\mathcal{P}_1, \ldots, \mathcal{P}_{i-1}, \mathcal{P}_{i+1}, \ldots, \mathcal{P}_n\}$ denote the set of players other than player $\mathcal{P}_i$.

2.2 Forms of Equilibrium

In this section we will introduce three forms of equilibrium that will be discussed in this dissertation: Nash equilibrium, correlated equilibrium (CE), and coarse correlated equilibrium (CCE).

2.2.1 Nash Equilibrium

The most well known form of an equilibrium is the Nash equilibrium.

Definition 2.2.1 (Pure Nash Equilibrium). *An action profile $a^* \in \mathcal{A}$ is called a pure Nash equilibrium if for all players $\mathcal{P}_i \in \mathcal{P}$,*

$$U_i(a_i^*, a_{-i}^*) = \max_{a_i \in \mathcal{A}_i} U_i(a_i, a_{-i}^*). \tag{2.1}$$

Furthermore, if the above condition is satisfied with a unique maximizer for every player $\mathcal{P}_i \in \mathcal{P}$, then a^* is called a strict Nash equilibrium.

A Nash equilibrium represents a scenario for which no player has an incentive to unilaterally deviate.

The concept of Nash equilibrium also extends to mixed strategy spaces. Let the *strategy* of player $\mathcal{P}_i$ be defined as $p_i \in \Delta(\mathcal{A}_i)$, where $\Delta(\mathcal{A}_i)$ is the set of probability distributions over the finite set of actions $\mathcal{A}_i$. We will adopt the convention that $p_i^{a_i}$ represents the probability that player $\mathcal{P}_i$ will select action a_i and $\sum_{a_i \in \mathcal{A}_i} p_i^{a_i} = 1$. If all players $\mathcal{P}_i \in \mathcal{P}$ play independently according to their personal strategy $p_i \in \Delta(\mathcal{A}_i)$, then the expected utility of player $\mathcal{P}_i$ for strategy p_i is defined as

$$U_i(p_i, p_{-i}) = \sum_{a \in \mathcal{A}} U_i(a) p_1^{a_1} p_2^{a_2} \dots p_n^{a_n},$$

where $p_{-i} = \{p_1, \dots, p_{i-1}, p_{i+1}, \dots, p_n\}$ denotes the collection of strategies of players other than player $\mathcal{P}_i$.

Definition 2.2.2 (Nash Equilibrium). *A strategy profile $p^* = \{p_1^*, \dots, p_n^*\}$ is called a Nash equilibrium if for all players $\mathcal{P}_i \in \mathcal{P}$,*

$$U_i(p_i^*, p_{-i}^*) = \max_{p_i \in \Delta(\mathcal{A}_i)} U_i(p_i, p_{-i}^*). \tag{2.2}$$

2.2.2 Correlated Equilibrium

In this section we will define a broader class of equilibria for which there may be correlations among the players. To that end, let $z \in \Delta(\mathcal{A})$ denote a probability distribution over the set of joint actions $\mathcal{A}$. We will adopt the convention that z^a is the probability of the joint action a and $\sum_{a \in \mathcal{A}} z^a = 1$. In the special case that all players $\mathcal{P}_i \in \mathcal{P}$ play independently according to their personal strategy $p_i \in \Delta(\mathcal{A}_i)$, as was the case in the definition of the Nash equilibrium, then

$$z^a = p_1^{a_1} p_2^{a_2} \dots p_n^{a_n},$$

where $a = (a_1, a_2, \dots, a_n)$.

Definition 2.2.3 (Correlated Equilibrium). *The probability distribution z is a correlated equilibrium if for all players $\mathcal{P}_i \in \mathcal{P}$ and for all actions $a_i, a_i' \in \mathcal{A}_i$,*

$$\sum_{a_{-i} \in \mathcal{A}_{-i}} U_i(a_i, a_{-i}) z^{(a_i, a_{-i})} \geq \sum_{a_{-i} \in \mathcal{A}_{-i}} U_i(a_i', a_{-i}) z^{(a_i, a_{-i})}. \tag{2.3}$$

To motivate this definition consider the following scenario. First, a joint action $a \in \mathcal{A}$ is randomly drawn according to the probability distribution $z \in \Delta(\mathcal{A})$. Next, each player is informed of only his particular action a_i, but not the actions of the other players. Finally, each player is given the opportunity to change his action. The condition for correlated equilibrium in (2.3) states that each player $\mathcal{P}_i$'s conditional expected payoff for action a_i is at least as good as his conditional expected payoff for any other action $a_i' \neq a_i$. In other words, a probability distribution z is a correlated equilibrium if and only if no player would seek to change their action from the outcome, randomly drawn according to z, even after his part has been revealed.

Notice that all Nash equilibria are in fact correlated equilibria.

2.2.3 Coarse Correlated Equilibrium

We will now relax the requirements on correlated equilibrium. Before doing so, we will discuss marginal distributions. Given the joint distribution $z \in \Delta(\mathcal{A})$, the marginal distribution of all players other than player $\mathcal{P}_i$ is

$$z_{-i}^{a_{-i}} = \sum_{a_i' \in \mathcal{A}_i} z^{(a_i', a_{-i})}.$$

Note that z_{-i} is a well defined probability distribution in $\Delta(\mathcal{A}_{-i})$.

Definition 2.2.4 (Coarse Correlated Equilibrium). *The probability distribution z is a coarse correlated equilibrium if for all players $\mathcal{P}_i \in \mathcal{P}$ and for all actions $a_i' \in \mathcal{A}_i$,*

$$\sum_{a \in \mathcal{A}} U_i(a) z^a \geq \sum_{a_{-i} \in \mathcal{A}_{-i}} U_i(a_i', a_{-i}) z_{-i}^{a_{-i}}. \tag{2.4}$$

To motivate this definition, consider the following scenario which differs slightly from the correlated equilibrium scenario. Before the joint action a is drawn, each player $\mathcal{P}_i$ is given the opportunity to opt out, in which case the player can select any action $a_i \in \mathcal{A}_i$ that he wishes. If the player does not opt out, he commits himself to playing his part of the action-tuple a randomly drawn according to the distribution z. In words, a distribution z is a coarse correlated equilibrium if under this scenario no player would choose to opt out given that all other players opt to stay in.

If the joint distribution z is a correlated equilibrium, then we know that for any action $a_i' \in \mathcal{A}_i$

$$\begin{aligned}
\sum_{a_i \in \mathcal{A}_i} \sum_{a_{-i} \in \mathcal{A}_{-i}} U_i(a_i, a_{-i}) z^{(a_i, a_{-i})} &\geq \sum_{a_i \in \mathcal{A}_i} \sum_{a_{-i} \in \mathcal{A}_{-i}} U_i(a_i', a_{-i}) z^{(a_i, a_{-i})}, \\
&= \sum_{a_{-i} \in \mathcal{A}_{-i}} U_i(a_i', a_{-i}) \sum_{a_i \in \mathcal{A}_i} z^{(a_i, a_{-i})}, \\
&= \sum_{a_{-i} \in \mathcal{A}_{-i}} U_i(a_i', a_{-i}) z_{-i}^{a_{-i}}.
\end{aligned}$$

This implies that for any action $a_i' \in \mathcal{A}_i$

$$\sum_{a \in \mathcal{A}} U_i(a) z^a \geq \sum_{a_{-i} \in \mathcal{A}_{-i}} U_i(a_i', a_{-i}) z_{-i}^{a_{-i}}.$$

Therefore, all correlated equilibria, and hence Nash equilibria, are in fact coarse correlated equilibria as illustrated in Figure 2.2. Under the condition that all players select their action independently, as was the case in the definition of the Nash equilibrium, then the definition of correlated, coarse correlated, and Nash equilibria are all equivalent.

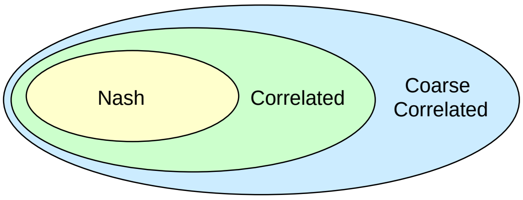


Figure 2.2: Relationship Between Nash, Correlated, and Coarse Correlated Equilibria.

2.2.4 Equilibrium Comparison

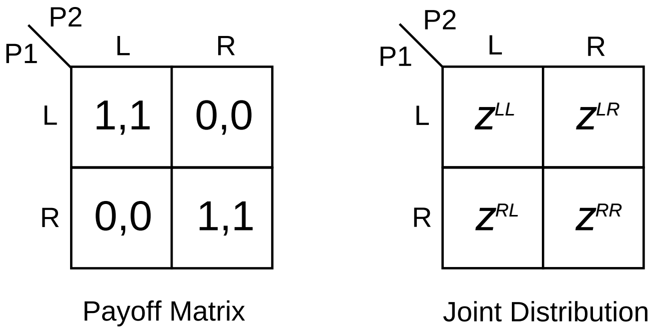
The main difference between Nash, correlated, and coarse correlated equilibria is whether a player is committed conditionally or unconditionally to a random draw of a given joint distribution $z \in \Delta(\mathcal{A})$. Table 2.1, taken from [You05], summarizes the main differences between the three forms of equilibria.

	Conditional Participation	Unconditional Participation
Independent Probabilities	Nash	Nash
Correlated Probabilities	Correlated	Coarse Correlated

Table 2.1: Relationship Between Nash, Correlated, and Coarse Correlated Equilibria.

We will now present a simple two player example, from [You05], to highlight the differences between the set of Nash equilibria and the set of correlated or coarse correlated equilibria. Note that the set of correlated equilibria and the set of coarse correlated equilibria are equivalent in two player games.

Consider the following two player game with payoff matrix as illustrated if Figure 2.3. For any joint action, the first entry is the payoff for player 1 and the second entry is the payoff for player 2. For example, $U_1(L, L) = 1$ and $U_2(L, L) = 1$. Let $z = \{z^{LL}, z^{LR}, z^{RL}, z^{LL}\}$ be a probability distribution over the joint action space $\mathcal{A} = \{LL, LR, RL, RR\}$.



P1 \ P2	L	R
L	1,1	0,0
R	0,0	1,1

Payoff Matrix

P1 \ P2	L	R
L	z^{LL}	z^{LR}
R	z^{RL}	z^{RR}

Joint Distribution

Figure 2.3: Example of an Identical Interest Game

In this example, there are two strict Nash equilibria, (L, L) and (R, R). Furthermore, there is one mixed Nash equilibrium, $p_1^L = p_2^L = 1/2$ and $p_1^R = p_2^R = 1/2$. A joint distribution z is a correlated equilibrium if and only if the off-diagonal probabilities do not exceed the diagonal probabilities, i.e.,

$$\max\{z^{LR}, z^{RL}\} \leq \min\{z^{LL}, z^{RR}\}.$$

Therefore, the set of correlated equilibria is significantly larger than the set of Nash equilibria.

2.3 Classes of Games

In this dissertation we will consider four classes of games: identical interest games, potential games, congestion games, and weakly acyclic games. Each class of games imposes a restriction on the admissible utility functions.

2.3.1 Identical Interest Games

The most restrictive class of games that we will review in this dissertation is identical interest games. In such a game, the players' utility functions $\{U_i\}_{i=1}^n$ are chosen to be the same. That is, for some function $\phi : \mathcal{A} \to \mathbb{R}$,

$$U_i(a) = \phi(a),$$

for every $\mathcal{P}_i \in \mathcal{P}$ and for every $a \in \mathcal{A}$. It is easy to verify that all identical interest games have at least one pure Nash equilibrium, namely any action profile a that maximizes $\phi(a)$. An example of an identical interest game is illustrated in Figure 2.3.

2.3.2 Potential Games

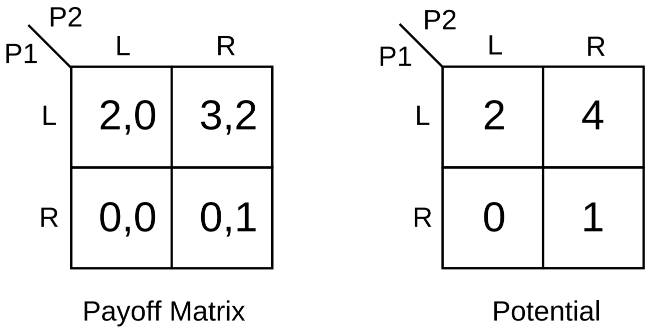
A significant generalization of an identical interest game is a potential game. In a potential game, the change in a player's utility that results from a unilateral change in strategy equals the change in the global utility. Specifically, there is a function $\phi : \mathcal{A} \to \mathbb{R}$ such that for every player $\mathcal{P}_i \in \mathcal{P}$, for every $a_{-i} \in \mathcal{A}_{-i}$, and for every $a_i', a_i'' \in \mathcal{A}_i$,

$$U_i(a_i', a_{-i}) - U_i(a_i'', a_{-i}) = \phi(a_i', a_{-i}) - \phi(a_i'', a_{-i}). \tag{2.5}$$

When this condition is satisfied, the game is called a potential game with the potential function ϕ. It is easy to see that in potential games, any action profile maximizing the potential function is a pure Nash equilibrium, hence every potential game possesses at least one such equilibrium.

An example of a two player potential game with associated potential function is illustrated if Figure 2.4.

We will also consider a more general class of potential games known as *generalized ordinal potential games*. In generalized ordinal potential games there is a function



P1 \ P2	L	R
L	2,0	3,2
R	0,0	0,1

Payoff Matrix

P1 \ P2	L	R
L	2	4
R	0	1

Potential

Figure 2.4: Example of a Potential Game with Potential Function

$\phi : \mathcal{A} \rightarrow \mathbb{R}$ such that for every player $\mathcal{P}_i \in \mathcal{P}$, for every $a_{-i} \in \mathcal{A}_{-i}$, and for every $a_i', a_i'' \in \mathcal{A}_i$,

$$U_i(a_i', a_{-i}) - U_i(a_i'', a_{-i}) > 0 \;\Rightarrow\; \phi(a_i', a_{-i}) - \phi(a_i'', a_{-i}) > 0.$$

2.3.3 Congestion Games

Congestion games are a specific class of games in which player utility functions have a special structure.

In order to define a congestion game, we must specify the action set, $\mathcal{A}_i$, and utility function, $U_i(\cdot)$, of each player. Towards this end, let $\mathcal{R}$ denote a finite set of "resources". For each resource $r \in \mathcal{R}$, there is an associated "congestion function"

$$c_r : \{0, 1, 2, ...\} \rightarrow \mathbb{R}$$

that reflects the cost of using the resource as a function of the number of players using that resource.

The action set, $\mathcal{A}_i$, of each player, $\mathcal{P}_i$, is defined as the set of resources available to player $\mathcal{P}_i$, i.e.,

$$\mathcal{A}_i \subset 2^{\mathcal{R}},$$

where $2^{\mathcal{R}}$ denotes the set of subsets of $\mathcal{R}$. Accordingly, an action, $a_i \in \mathcal{A}_i$, reflects a selection of (multiple) resources, $a_i \subset \mathcal{R}$. A player is "using" resource r if $r \in a_i$. For an action profile $a \in \mathcal{A}$, let $\sigma_r(a)$ denote the total number of players using resource r, i.e., $|\{i : r \in a_i\}|$. In a congestion game, the utility of player $\mathcal{P}_i$ using resources indicated by a_i depends only on the total number of players using the same resources. More precisely, the utility of player $\mathcal{P}_i$ is defined as

$$U_i(a) = \sum_{r \in a_i} c_r(\sigma_r(a)). \tag{2.6}$$

Any congestion game with utility functions as in (2.6) is a potential game [Ros73] with potential function

$$\phi(a) = \sum_{r \in \mathcal{R}} \sum_{k=1}^{\sigma_r(a)} c_r(k). \tag{2.7}$$

In fact, every congestion game is a potential game and every finite potential game is isomorphic to a congestion game [MS96b].

2.3.4 Weakly Acyclic Games

Consider any finite game G with a set $\mathcal{A}$ of action profiles. A *better reply path* is a sequence of action profiles $a^1, a^2, ..., a^L$ such that, for every $1 \leq \ell \leq L-1$, there is exactly one player $\mathcal{P}_{i_\ell}$ such that i) $a^{\ell}_{i_\ell} \neq a^{\ell+1}_{i_\ell}$, ii) $a^{\ell}_{-i_\ell} = a^{\ell+1}_{-i_\ell}$, and iii) $U_{i_\ell}(a^\ell) < U_{i_\ell}(a^{\ell+1})$. In other words, one player moves at a time, and each time a player moves he increases his own utility.

Suppose now that G is a potential game with potential function ϕ. Starting from an arbitrary action profile $a \in \mathcal{A}$, construct a better reply path $a = a^1, a^2, ..., a^L$ until it can no longer be extended. Note first that such a path cannot cycle back on itself, because ϕ is strictly increasing along the path. Since $\mathcal{A}$ is finite, the path cannot be extended indefinitely. Hence, the last element in a maximal better reply path from any joint action, a, must be a Nash equilibrium of G.

This idea may be generalized as follows. The game G is *weakly acyclic* if for any $a \in \mathcal{A}$, there exists a better reply path starting at a and ending at some pure Nash equilibrium of G [You98, You05]. Potential games are special cases of weakly acyclic games.

An example of a two player weakly acyclic game is illustrated in Figure 2.5.

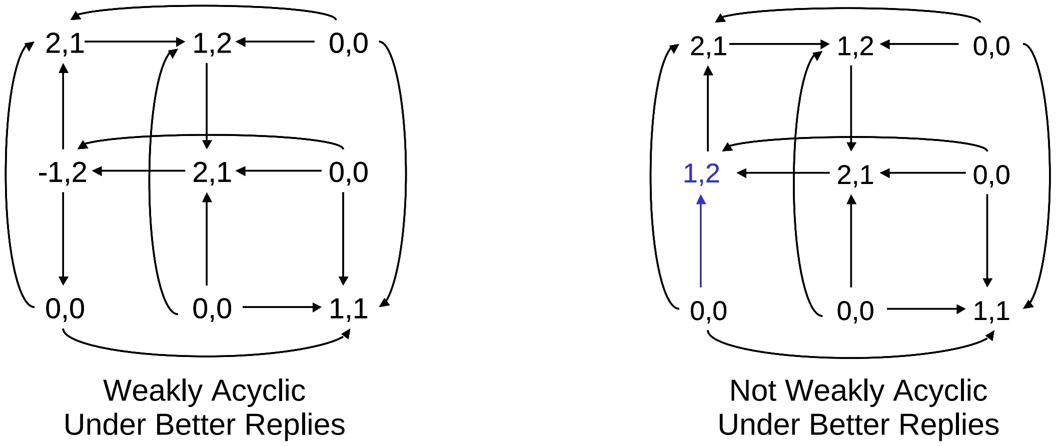


Figure 2.5: Example of a Weakly Acyclic Game

2.4 Repeated Games

In a repeated game, at each time $t \in \{0, 1, 2, \dots\}$, each player $\mathcal{P}_i \in \mathcal{P}$ simultaneously chooses an action $a_i(t) \in \mathcal{A}_i$ and receives the utility $U_i(a(t))$ where $a(t) := (a_1(t), \dots, a_n(t))$. Each player $\mathcal{P}_i \in \mathcal{P}$ chooses his action $a_i(t)$ at time t simultaneously according to a probability distribution $p_i(t)$, which we will refer to as the *strategy* of player $\mathcal{P}_i$ at time t. A player's strategy at time t can rely only on observations from times $\{0, 1, 2, ..., t-1\}$. Different learning algorithms are specified by both the assumptions on available information and the mechanism by which the strategies are updated as information is gathered.

We will review three main classes of learning algorithms in this dissertation: full information, virtual payoff based, and payoff based. For a detailed review of learning in games we direct the reader to [FL98, You98, You05, HS98, Wei95, Sam97].

2.4.1 Full Information Learning Algorithms

The most informationally sophisticated class of learning algorithms is full information. In full information learning algorithms, each player knows the functional form of his utility function and is capable of observing the actions of all other players at every time step. The strategy adjustment mechanism of player $\mathcal{P}_i$ can be written in the general form

$$p_i(t) = F_i\big(a(0), ..., a(t-1); U_i\big).$$

In this setting, players may develop probabilistic models for the actions of other players using past observations. Based off these models, players may seek to maximize some form of an expected utility. An example of a learning algorithm, or strategy adjustment mechanism, of this form is the well known fictitious play [MS96a]. We will review fictitious play in Section 3.2.1.

2.4.2 Virtual Payoff Based Learning Algorithms

We will now relax the requirements of full information learning algorithms. In virtual payoff based algorithms, players are now unaware of the structural form of their utility function. Furthermore, players also are not capable of observing the actions of all players. However, players are endowed with the ability to assess the utility that they would have received for alternative action choices. For example, suppose that the action played at time t is $a(t)$. In virtual payoff based dynamics, each player $\mathcal{P}_i$ with action set $\mathcal{A}_i = \{a_i^1, ..., a_i^{|\mathcal{A}_i|}\}$ has access to the following information:

$$a(t) \Rightarrow \begin{bmatrix} U_i(a_i^1, a_{-i}(t)) \\ \vdots \\ U_i(a_i^{|\mathcal{A}_i|}, a_{-i}(t)) \end{bmatrix},$$

where $|\mathcal{A}_i|$ denotes the cardinality of the action set $\mathcal{A}_i$.

The strategy adjustment mechanism of player $\mathcal{P}_i$ can be written in the general form

$$p_i(t) = F_i\big(\ \{U_i(a_i, a_{-i}(0))\}_{a_i \in \mathcal{A}_i},\ \ldots,\ \{U_i(a_i, a_{-i}(t-1))\}_{a_i \in \mathcal{A}_i}\ \big).$$

An example of a learning algorithm, or strategy adjustment mechanism, of this form is the well known regret matching [HM00]. We will review regret matching in Section 4.2. Virtual payoff based learning algorithms will be the focus of Chapters 3 and 4.

2.4.3 Payoff Based Learning Algorithms

Payoff based learning algorithms are the most informationally restrictive class of learning algorithms. Now, players *only* have access to (i) the action they played and (ii) the utility (possibly noisy) they received. In this setting, the strategy adjustment mechanism of player $\mathcal{P}_i$ takes on the form

$$p_i(t) = F_i\big(\{a_i(0), U_i(a(0))\}, ..., \{a_i(t-1), U_i(a(t-1))\}\big). \tag{2.8}$$

We will discuss payoff based learning algorithms extensively in Chapter 5.

CHAPTER 3

Joint Strategy Fictitious Play with Inertia for Potential Games

In this chapter we consider multi-player repeated games involving a large number of players with large strategy spaces and enmeshed utility structures. In these "large-scale" games, players are inherently faced with limitations in both their observational and computational capabilities. Accordingly, players in large-scale games need to make their decisions using algorithms that accommodate limitations in information gathering and processing. This disqualifies some of the well known decision making models such as "Fictitious Play" (FP), in which each player must monitor the individual actions of every other player and must optimize over a high dimensional probability space. We will show that Joint Strategy Fictitious Play (JSFP), a close variant of FP, alleviates both the informational and computational burden of FP. Furthermore, we introduce JSFP with inertia, i.e., a probabilistic reluctance to change strategies, and establish the convergence to a pure Nash equilibrium in all generalized ordinal potential games in both cases of averaged or exponentially discounted historical data. We illustrate JSFP with inertia on the specific class of congestion games, a subset of generalized ordinal potential games. In particular, we illustrate the main results on a distributed traffic routing problem and derive tolling procedures that can lead to optimized total traffic congestion.

3.1 Introduction

We consider "large-scale" repeated games involving a large number of players, each of whom selects a strategy from a possibly large strategy set. A player's reward, or utility, depends on the actions taken by all players. The game is repeated over multiple stages, and this allows players to adapt their strategies in response to the available information gathered over prior stages. This setup falls under the general subject of "learning in games" [FL98, You05], and there are a variety of algorithms and accompanying analysis that examine the long term behavior of these algorithms.

In large-scale games players are inherently faced with limitations in both their observational and computational capabilities. Accordingly, players in such large-scale games need to make their decisions using algorithms that accommodate limitations in information gathering and processing. This limits the feasibility of different learning algorithms. For example, the well-studied algorithm "Fictitious Play" (FP) requires individual players to individually monitor the actions of other players and to optimize their strategies according to a probability distribution function over the joint actions of other players. Clearly, such information gathering and processing is not feasible in a large-scale game.

The main objective of this chapter is to study a variant of FP called Joint Strategy Fictitious Play (JSFP) [FL98, FK93, MS97]. We will argue that JSFP is a plausible decision making model for certain large-scale games. We will introduce a modification of JSFP to include inertia, in which there is a probabilistic reluctance of any player to change strategies. We will establish that JSFP with inertia converges to a pure Nash equilibrium for a class of games known as generalized ordinal potential games, which includes so-called congestion games as a special case [Ros73].

Our motivating example for a large-scale congestion game is distributed traffic routing [BL85], in which a large number of vehicles make daily routing decisions to optimize their own objectives in response to their own observations. In this setting, observing and responding to the individual actions of all vehicles on a daily basis would be a formidable task for any individual driver. A more realistic measurement on the information tracked and processed by an individual driver is the daily aggregate congestion on the roads that are of interest to that driver [BPK91]. It turns out that JSFP accommodates such information aggregation.

We will now review some of the well known decision making models and discuss their limitations in large-scale games. See the monographs [FL98, You98, You05, HS98, Wei95] and survey article [Har05] for a more comprehensive review.

The well known FP algorithm requires that each player views all other players as independent decision makers [FL98]. In the FP framework, each player observes the decisions made by all other players and computes the empirical frequencies (i.e. running averages) of these observed decisions. Then, each player best responds to the empirical frequencies of other players' decisions by first computing the expected utility for each strategy choice under the assumption that the other players will independently make their decisions probabilistically according to the observed empirical frequencies. FP is known to be convergent to a Nash equilibrium in potential games, but need not converge for other classes of games. General convergence issues are discussed in [HM03b, SA05, AS04].

The paper [LES05] introduces a version of FP, called "sampled FP", that seeks to avoid computing an expected utility based on the empirical frequencies, because for large scale games, this expected utility computation can be prohibitively demanding. In sampled FP, each player selects samples from the strategy space of every other player according to the empirical frequencies of that player's past decisions. A player then computes an average utility for each strategy choice based off of these samples.

Each player still has to observe the decisions made by all other players to compute the empirical frequencies of these observed decisions. Sampled FP is proved to be convergent in identical interest games, but the number of samples needed to guarantee convergence grows unboundedly.

There are convergent learning algorithms for a large class of coordination games called "weakly acyclic" games [You98]. In adaptive play [You93] players have finite recall and respond to the recent history of other players. Adaptive play requires each player to track the individual behavior of all other players for recall window lengths greater than one. Thus, as the size of player memory grows, adaptive play suffers from the same computational setback as FP.

It turns out that there is a strong similarity between the JSFP discussed herein and the regret matching algorithm [HM00]. A player's regret for a particular choice is defined as the difference between 1) the utility that would have been received if that particular choice was played for all the previous stages and 2) the average utility actually received in the previous stages. A player using the regret matching algorithm updates a regret vector for each possible choice, and selects actions according to a probability proportional to positive regret. In JSFP, a player chooses an action by myopically maximizing the anticipated utility based on past observations, which is effectively equivalent to regret modulo a bias term. A current open question is whether player choices would converge in coordination-type games when all players use the regret matching algorithm (except for the special case of two-player games [HM03a]). There are finite memory versions of the regret matching algorithm and various generalizations [You05], such as playing best or better responses to regret over the last m stages, that are proven to be convergent in weakly acyclic games when players use some sort of inertia. These finite memory algorithms do not require each player to track the behavior of other players individually. Rather, each player needs to remember the utilities actually received and the utilities that could have been received in the last m stages. In contrast, a player using JSFP best responds according to accumulated experience over the entire history by using a simple recursion which can also incorporate exponential discounting of the historical data.

There are also payoff based dynamics, where each player observes only the actual utilities received and uses a Reinforcement Learning (RL) algorithm [SB98, BT96] to make future choices. Convergence of player choices when all players use an RL-like algorithm is proved for identical interest games [LC03, LC05b, LC05a] assuming that learning takes place at multiple time scales. Finally, the payoff based dynamics with finite-memory presented in [HS04] leads to a Pareto-optimal outcome in generic common interest games.

Regarding the distributed routing setting of Section 3.4, there are papers that analyze different routing strategies in congestion games with "infinitesimal" players, i.e., a continuum of players as opposed to a large, but finite, number of players. References [FV04, FV05, FRV06] analyze the convergence properties of a class of routing strategies that is a variation of the replicator dynamics in congestion games, also referred to as symmetric games, under a variety of settings. Reference [BEL06] analyzes the convergence properties of no-regret algorithms in such congestion games and also considers congestion games with discrete players, as considered in this paper, but the results hold only for a highly structured symmetric game.

The remainder of this chapter is organized as follows. Section 3.2, sets up JSFP and goes on to establish convergence to a pure Nash equilibrium for JSFP with inertia in all generalized ordinal potential games. Section 3.3 presents a fading memory variant of JSFP, and likewise establishes convergence to a pure Nash equilibrium. Section 3.4 presents an illustrative example for traffic congestion games. Section 3.4 goes on to illustrate the use of tolls to achieve a socially optimal equilibrium and derives conditions for this equilibrium to be unique.

3.2 Joint Strategy Fictitious Play with Inertia

Consider a finite game with n-player set $\mathcal{P} := \{\mathcal{P}_1, ..., \mathcal{P}_n\}$ where each player $\mathcal{P}_i \in \mathcal{P}$ has an action set $\mathcal{A}_i$ and a utility function $U_i : \mathcal{A} \rightarrow \mathbb{R}$ where $\mathcal{A} = \mathcal{A}_1 \times ... \times \mathcal{A}_n$.

In a *repeated game* as described in Section 2.4, at every stage $t \in \{0, 1, 2, ...\}$, each player, $\mathcal{P}_i$, simultaneously selects an action $a_i(t) \in \mathcal{A}_i$. This selection is a function of the information available to player $\mathcal{P}_i$ up to stage t. Both the action selection function and the available information depend on the underlying learning process.

3.2.1 Fictitious Play

We start with the well known Fictitious Play (FP) process [FL98]. Fictitious Play is an example of a full information learning algorithm.

Define the *empirical frequency*, $q_i^{\bar{a}_i}(t)$, as the percentage of stages at which player $\mathcal{P}_i$ has chosen the action $\bar{a}_i \in \mathcal{A}_i$ up to time $t-1$, i.e.,

$$q_i^{\bar{a}_i}(t) := \frac{1}{t} \sum_{\tau=0}^{t-1} I\{a_i(\tau) = \bar{a}_i\},$$

where $a_i(k) \in \mathcal{A}_i$ is player $\mathcal{P}_i$'s action at time k and $I\{\cdot\}$ is the indicator function. Now define the empirical frequency vector for player $\mathcal{P}_i$ as

$$q_i(t) := \begin{pmatrix} q_i^{\bar{a}_1} \\ \vdots \\ q_i^{\bar{a}_{|\mathcal{A}_i|}} \end{pmatrix},$$

where $|\mathcal{A}_i|$ is the cardinality of the action set $\mathcal{A}_i$.

The action of player $\mathcal{P}_i$ at time t is based on the (incorrect) presumption that other players are playing *randomly* and *independently* according to their empirical frequencies. Under this presumption, the expected utility for the action $\bar{a}_i \in \mathcal{A}_i$ is

$$U_i(\bar{a}_i, q_{-i}(t)) := \sum_{a_{-i} \in \mathcal{A}_{-i}} U_i(\bar{a}_i, a_{-i}) \prod_{a_j \in a_{-i}} q_j^{a_j}(t), \tag{3.1}$$

where $q_{-i}(t) := \{q_1(t), ..., q_{i-1}(t), q_{i+1}(t), ..., q_n(t)\}$ and $\mathcal{A}_{-i} := \times_{j \neq i} \mathcal{A}_j$. In the FP process, player $\mathcal{P}_i$ uses this expected utility by selecting an action at time t from the set

$$BR_i(q_{-i}(t)) := \{\tilde{a}_i \in \mathcal{A}_i : U_i(\tilde{a}_i, q_{-i}(t)) = \max_{a_i \in \mathcal{A}_i} U_i(a_i, q_{-i}(t))\}.$$

The set $BR_i(q_{-i}(t))$ is called player $\mathcal{P}_i$'s best response to $q_{-i}(t)$. In case of a non-unique best response, player $\mathcal{P}_i$ makes a random selection from $BR_i(q_{-i}(t))$.

It is known that the empirical frequencies generated by FP converge to a Nash equilibrium in potential games [MS96b].

Note that FP as described above requires each player to observe the actions made by every other individual player. Moreover, choosing an action based on the predictions (3.1) amounts to enumerating all possible joint actions in $\times_j \mathcal{A}_j$ at every stage for each player. Hence, FP is computationally prohibitive as a decision making model in large-scale games.

3.2.2 Setup: Joint Strategy Fictitious Play

In JSFP, each player tracks the empirical frequencies of the *joint actions* of all other players. In contrast to FP, the action of player $\mathcal{P}_i$ at time t is based on the (still incorrect) presumption that other players are playing *randomly* but *jointly* according to their *joint* empirical frequencies, i.e., each player views all other players as a collective group.

Let $z^{\bar{a}}(t)$ be the percentage of stages at which all players chose the joint action profile $\overline{a} \in \mathcal{A}$ up to time $t-1$, i.e.,

$$z^{\bar{a}}(t) := \frac{1}{t}\sum_{\tau=0}^{t-1} I\{a(\tau) = \bar{a}\}. \tag{3.2}$$

Let $z(t)$ denote the empirical frequency vector formed by the components $\{z^{\bar{a}}(t)\}_{\bar{a}\in\mathcal{A}}$. Note that the dimension of $z(t)$ is the cardinality $|\mathcal{A}|$.

Similarly, let $z_{-i}^{\overline{a}_{-i}}(t)$ be the percentage of stages at which players other then player $\mathcal{P}_i$ have chosen the joint action profile $\overline{a}_{-i} \in \mathcal{A}_{-i}$ up to time $t-1$, i.e.,

$$z_{-i}^{\bar{a}_{-i}}(t) := \frac{1}{t}\sum_{\tau=0}^{t-1} I\{a_{-i}(\tau) = \bar{a}_{-i}\}, \tag{3.3}$$

which, given $z(t)$, can also be expressed as

$$z_{-i}^{\bar{a}_{-i}}(t) = \sum_{a_i\in\mathcal{A}_i} z^{(a_i,\bar{a}_{-i})}(t).$$

Let $z_{-i}(t)$ denote the empirical frequency vector formed by the components $\{z_{-i}^{\bar{a}_{-i}}(t)\}_{\bar{a}_{-i}\in\mathcal{A}_{-i}}$. Note that the dimension of $z_{-i}(t)$ is the cardinality $|\times_{i\neq j}\mathcal{A}_j|$.

Similarly to FP, player $\mathcal{P}_i$'s action at time t is based on an expected utility for the action $\bar{a}_i \in \mathcal{A}_i$, but now based on the joint action model of opponents given by[1]

$$U_i(\bar{a}_i, z_{-i}(t)) := \sum_{a_{-i}\in\mathcal{A}_{-i}} U_i(\bar{a}_i, a_{-i}) z_{-i}^{a_{-i}}(t). \tag{3.4}$$

In the JSFP process, player $\mathcal{P}_i$ uses this expected utility by selecting an action at time t from the set

$$BR_i(z_{-i}(t)) := \{\tilde{a}_i \in \mathcal{A}_i : U_i(\tilde{a}_i, z_{-i}(t)) = \max_{a_i\in\mathcal{A}_i} U_i(a_i, z_{-i}(t))\}.$$

Note that the utility as expressed in (3.4) is linear in $z_{-i}(t)$.

When written in this form, JSFP appears to have a computational burden for each player that is even higher than that of FP, since tracking the empirical frequencies

[1] Note that we use the same notation for the related quantities $U(a_i, a_{-i})$, $U(a_i, q_{-i})$, and $U(a_i, z_{-i})$, where the latter two are derived from the first as defined in equations (3.1) and (3.4), respectively.

$z_{-i}(t) \in \Delta(\mathcal{A}_{-i})$ of the joint actions of the other players is more demanding for player $\mathcal{P}_i$ than tracking the empirical frequencies $q_{-i}(t) \in \times_{j \neq i} \Delta(\mathcal{A}_j)$ of the actions of the other players individually, where $\Delta(\mathcal{A})$ denotes the set of probability distributions on a finite set $\mathcal{A}$. However, it is possible to rewrite JSFP to significantly reduce the computational burden on each player.

To choose an action at any time, t, player $\mathcal{P}_i$ using JSFP needs only the predicted utilities $U_i(\bar{a}_i, z_{-i}(t))$ for each $\bar{a}_i \in \mathcal{A}_i$. Substituting (3.3) into (3.4) results in

$$U_i(\bar{a}_i, z_{-i}(t)) = \frac{1}{t} \sum_{\tau=0}^{t-1} U_i(\bar{a}_i, a_{-i}(\tau)),$$

which is the average utility player $\mathcal{P}_i$ would have received if action $\bar{a}_i$ had been chosen at every stage up to time $t-1$ and other players used the same actions. This average utility, denoted by $V_i^{\bar{a}_i}(t)$, admits the following simple recursion,

$$V_i^{\bar{a}_i}(t+1) = \frac{t}{t+1} V_i^{\bar{a}_i}(t) + \frac{1}{t+1} U_i(\bar{a}_i, a_{-i}(t)).$$

The important implication is that JSFP dynamics can be implemented *without* requiring each player to track the empirical frequencies of the joint actions of the other players and *without* requiring each player to compute an expectation over the space of the joint actions of all other players. Rather, each player using JSFP merely updates the predicted utilities for each available action using the recursion above, and chooses an action each stage with maximal predicted utility.

An interesting feature of JSFP is that each strict Nash equilibrium has an “absorption” property as summarized in Proposition 3.2.1.

Proposition 3.2.1. *In any finite n-person game, if at any time $t > 0$, the joint action $a(t)$ generated by a JSFP process is a strict Nash equilibrium, then $a(t+\tau) = a(t)$ for all $\tau > 0$.*

Proof. For each player $\mathcal{P}_i \in \mathcal{P}$ and for all actions $a_i \in \mathcal{A}_i$,

$$U_i(a_i(t), z_{-i}(t)) \geq U_i(a_i, z_{-i}(t)).$$

Since $a(t)$ is a strict Nash equilibrium, we know that for all actions $a_i \in \mathcal{A}_i \backslash a_i(t)$

$$U_i(a_i(t), a_{-i}(t)) > U_i(a_i, a_{-i}(t)).$$

By writing $z_{-i}(t+1)$ in terms of $z_{-i}(t)$ and $a_{-i}(t)$,

$$U_i(a_i(t), z_{-i}(t+1)) = \frac{t}{t+1} U_i(a_i(t), z_{-i}(t)) + \frac{1}{t+1} U_i(a_i(t), a_{-i}(t)).$$

Therefore, $a_i(t)$ is the only best response to $z_{-i}(t+1)$,

$$U_i(a_i(t), z_{-i}(t+1)) > U_i(a_i, z_{-i}(t+1)), \quad \forall a_i \in \mathcal{A}_i \backslash a_i(t).$$

□

A strict Nash equilibrium need *not* possess this absorption property in general for standard FP when there are more than two players.[2]

The convergence properties, even for potential games, of JSFP in the case of more than two players is unresolved.[3] We will establish convergence of JSFP in the case where players use some sort of inertia, i.e., players are reluctant to switch to a better action.

The **JSFP with inertia** process is defined as follows. Players choose their actions according to the following rules:

[2] To see this, consider the following 3 player identical interest game. For all $\mathcal{P}_i \in \mathcal{P}$, let $\mathcal{A}_i = \{a, b\}$. Let the utility be defined as follows: $U(a,b,a) = U(b,a,a) = 1, U(a,a,a) = U(b,b,a) = 0, U(a,a,b) = U(b,b,b) = 1, U(a,b,b) = -1, U(b,a,b) = -100$. Suppose the first action played is $a(1) = \{a,a,a\}$. In the FP process each player will seek to deviate in the ensuing stage, $a(2) = \{b,b,b\}$. The joint action $\{b,b,b\}$ is a strict Nash equilibrium. One can easily verify that the ensuing action in a FP process will be $a(3) = \{a,b,a\}$. Therefore, a strict Nash equilibrium is not absorbing in the FP process with more than 2 players.

[3] For two player games, JSFP and standard FP are equivalent, hence the convergence results for FP hold for JSFP.

JSFP–1: If the action $a_i(t-1)$ chosen by player $\mathcal{P}_i$ at time $t-1$ belongs to $BR_i(z_{-i}(t))$, then $a_i(t) = a_i(t-1)$.

JSFP–2: Otherwise, player $\mathcal{P}_i$ chooses an action, $a_i(t)$, at time t according to the probability distribution

$$\alpha_i(t)\beta_i(t) + (1 - \alpha_i(t))\mathbf{v}^{a_i(t-1)},$$

where $\alpha_i(t)$ is a parameter representing player $\mathcal{P}_i$'s willingness to optimize at time t, $\beta_i(t) \in \Delta(\mathcal{A}_i)$ is any probability distribution whose support is contained in the set $BR_i(z_{-i}(t))$, and $\mathbf{v}^{a_i(t-1)}$ is the probability distribution with full support on the action $a_i(t-1)$, i.e.,

$$\mathbf{v}^{a_i(t-1)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

where the "1" occurs in the coordinate of $\Delta(\mathcal{A}_i)$ associated with $a_i(t-1)$.

According to these rules, player $\mathcal{P}_i$ will stay with the previous action $a_i(t-1)$ with probability $1 - \alpha_i(t)$ even when there is a perceived opportunity for utility improvement. We make the following standing assumption on the players' willingness to optimize.

Assumption 3.2.1. *There exist constants $\underline{\varepsilon}$ and $\bar{\varepsilon}$ such that for all time $t \geq 0$ and for all players $\mathcal{P}_i \in \mathcal{P}$,*

$$0 < \underline{\varepsilon} < \alpha_i(t) < \bar{\varepsilon} < 1.$$

This assumption implies that players are always willing to optimize with some nonzero inertia[4].

The following result shows a similar absorption property of pure Nash equilibria in a JSFP with inertia process.

Proposition 3.2.2. *In any finite n-person game, if at any time $t > 0$ the joint action $a(t)$ generated by a JSFP with inertia process is 1) a pure Nash equilibrium and 2) the action $a_i(t) \in BR_i(z_{-i}(t))$ for all players $\mathcal{P}_i \in \mathcal{P}$, then $a(t+\tau) = a(t)$ for all $\tau > 0$.*

Proof. For each player $\mathcal{P}_i \in \mathcal{P}$ and for all actions $a_i \in \mathcal{A}_i$,

$$U_i(a_i(t), z_{-i}(t)) \geq U_i(a_i, z_{-i}(t)).$$

Since $a(t)$ is a pure Nash equilibrium, we know that for all actions $a_i \in \mathcal{A}_i$

$$U_i(a_i(t), a_{-i}(t)) \geq U_i(a_i, a_{-i}(t)).$$

By writing $z_{-i}(t+1)$ in terms of $z_{-i}(t)$ and $a_{-i}(t)$,

$$U_i(a_i(t), z_{-i}(t+1)) = \frac{t}{t+1} U_i(a_i(t), z_{-i}(t)) + \frac{1}{t+1} U_i(a_i(t), a_{-i}(t)).$$

Therefore, $a_i(t)$ is also a best response to $z_{-i}(t+1)$,

$$U_i(a_i(t), z_{-i}(t+1)) \geq U_i(a_i, z_{-i}(t+1)), \quad \forall a_i \in \mathcal{A}_i.$$

Since $a_i(t) \in BR_i(z_{-i}(t+1))$ for all players, then $a(t+1) = a(t)$. □

3.2.3 Convergence to Nash Equilibrium

The following establishes the main result regarding the convergence of JSFP with inertia.

We will assume that no player is indifferent between distinct strategies[5].

[4] This assumption can be relaxed to holding for sufficiently large t, as opposed to all t.

[5] One could alternatively assume that all pure equilibria are strict.

Assumption 3.2.2. *Player utilities satisfy*

$$U_i(a_i^1, a_{-i}) \neq U_i(a_i^2, a_{-i}),\ \forall\, a_i^1, a_i^2 \in \mathcal{A}_i,\ a_i^1 \neq a_i^2,\ \forall\, a_{-i} \in \mathcal{A}_{-i},\ \forall\, i \in \{1, ..., n\}. \tag{3.5}$$

Theorem 3.2.1. *In any finite generalized ordinal potential game in which no player is indifferent between distinct strategies as in Assumption 3.2.2, the action profiles* $a(t)$ *generated by JSFP with inertia under Assumption 3.2.1 converge to a pure Nash equilibrium almost surely.*

We provide a complete proof of Theorem 3.2.1 in the Appendix of this chapter. We encourage the reader to first review the proof of fading memory JSFP with inertia in Theorem 3.3.1 of the following section.

3.3 Fading Memory JSFP with Inertia

We now analyze the case where players view recent information as more important. In fading memory JSFP with inertia, players replace true empirical frequencies with weighted empirical frequencies defined by the recursion

$$\begin{aligned} \tilde{z}_{-i}^{\bar{a}_{-i}}(0) &:= I\{a_{-i}(0) = \bar{a}_{-i}\}, \\ \tilde{z}_{-i}^{\bar{a}_{-i}}(t) &:= (1-\rho)\tilde{z}_{-i}^{\bar{a}_{-i}}(t-1) + \rho I\{a_{-i}(t-1) = \bar{a}_{-i}\},\ \forall t \geq 1, \end{aligned}$$

where $0 < \rho \leq 1$ is a parameter with $(1-\rho)$ being the discount factor. Let $\tilde{z}_{-i}(t)$ denote the weighted empirical frequency vector formed by the components $\{\tilde{z}_{-i}^{\bar{a}_{-i}}(t)\}_{\bar{a}_{-i} \in \mathcal{A}_{-i}}$. Note that the dimension of $\tilde{z}_{-i}(t)$ is the cardinality $|\mathcal{A}_{-i}|$.

One can identify the limiting cases of the discount factor. When $\rho = 1$ we have “Cournot” beliefs, where only the most recent information matters. In the case when ρ is not a constant, but rather $\rho(t) = 1/(t+1)$, all past information is given equal importance as analyzed in Section 3.2.

Utility prediction and action selection with fading memory are done in the same way as in Section 3.2, and in particular, in accordance with rules JSFP-1 and JSFP-2. To make a decision, player $\mathcal{P}_i$ needs only the weighted average utility that would have been received for each action, which is defined for action $\bar{a}_i \in \mathcal{A}_i$ as

$$\tilde{V}_i^{\bar{a}_i}(t) := U_i(\bar{a}_i, \tilde{z}_{-i}(t)) = \sum_{a_{-i} \in \mathcal{A}_{-i}} U_i(\bar{a}_i, a_{-i}) \tilde{z}_{-i}^{a_{-i}}(t).$$

One can easily verify that the weighted average utility $\tilde{V}_i^{\bar{a}_i}(t)$ for action $\bar{a}_i \in \mathcal{A}_i$ admits the recursion

$$\tilde{V}_i^{\bar{a}_i}(t) = \rho U_i(\bar{a}_i, a_{-i}(t-1)) + (1-\rho)\tilde{V}_i^{\bar{a}_i}(t-1).$$

Once again, player $\mathcal{P}_i$ is not required to track the weighted empirical frequency vector $\tilde{z}_{-i}(t)$ or required to compute expectations over $\mathcal{A}_{-i}$.

As before, pure Nash equilibria have an absorption property under fading memory JSFP with inertia.

Proposition 3.3.1. *In any finite n-person game, if at any time $t > 0$ the joint action $a(t)$ generated by a fading memory JSFP with inertia process is 1) a pure Nash equilibrium and 2) the action $a_i(t) \in BR_i(\tilde{z}_{-i}(t))$ for all players $\mathcal{P}_i \in \mathcal{P}$, then $a(t+\tilde{t}) = a(t)$ for all $\tilde{t} > 0$.*

Proof. For each player $\mathcal{P}_i \in \mathcal{P}$ and for all actions $a_i \in \mathcal{A}_i$,

$$U_i(a_i(t), \tilde{z}_{-i}(t)) \geq U_i(a_i, \tilde{z}_{-i}(t)).$$

Since $a(t)$ is a pure Nash equilibrium, we know that for all actions $a_i \in \mathcal{A}_i$

$$U_i(a_i(t), a_{-i}(t)) \geq U_i(a_i, a_{-i}(t)).$$

By writing $\tilde{z}_{-i}(t+1)$ in terms of $\tilde{z}_{-i}(t)$ and $a_{-i}(t)$,

$$U_i(a_i(t), \tilde{z}_{-i}(t+1)) = (1-\rho)U_i(a_i(t), \tilde{z}_{-i}(t)) + \rho U_i(a_i(t), a_{-i}(t)).$$

Therefore, $a_i(t)$ is also a best response to $\tilde{z}_{-i}(t+1)$,

$$U_i(a_i(t), \tilde{z}_{-i}(t+1)) \geq U_i(a_i, \tilde{z}_{-i}(t+1)), \quad \forall a_i \in \mathcal{A}_i.$$

Since $a_i(t) \in BR_i(\tilde{z}_{-i}(t+1))$ for all players, then $a(t+1) = a(t)$. □

The following theorem establishes convergence to Nash equilibrium for fading memory JSFP with inertia.

Theorem 3.3.1. *In any finite generalized ordinal potential game in which no player is indifferent between distinct strategies as in Assumption 3.2.2, the action profiles $a(t)$ generated by a fading memory JSFP with inertia process satisfying Assumption 3.2.1 converge to a pure Nash equilibrium almost surely.*

Proof. The proof follows a similar structure to the proof of Theorem 6.2 in [You05]. At time t, let $a^0 := a(t)$. There exists a positive constant T, independent of t, such that if the current action a^0 is repeated T consecutive stages, i.e. $a(t) = ... = a(t+T-1) = a^0$, then $BR_i(\tilde{z}_{-i}(t+T)) = BR_i(a^0_{-i})$ [6] for all players. The probability of such an event is at least $(1-\bar{\varepsilon})^{n(T-1)}$, where n is the number of players. If the joint action a^0 is an equilibrium, then by Proposition 3.3.1 we are done. Otherwise, there must be at least one player $\mathcal{P}_{i(1)} \in \mathcal{P}$ such that $a^0_{i(1)} \notin BR_{i(1)}(a^0_{-i(1)})$ and hence $a^0_{i(1)} \notin BR_{i(1)}(\tilde{z}_{-i(1)}(t+T))$.

Consider now the event that, at time $t+T$, exactly one player switches to a different action, i.e., $a^1 := a(t+T) = (a^*_{i(1)}, a^0_{-i(1)})$ for some player $\mathcal{P}_{i(1)} \in \mathcal{P}$ where $U_{i(1)}(a^1) > U_{i(1)}(a^0)$. This event happens with probability at least $\underline{\varepsilon}(1-\bar{\varepsilon})^{n-1}$. Note that if $\phi(\cdot)$ is a generalized ordinal potential function for the game, then $\phi(a^0) < \phi(a^1)$.

Continuing along the same lines, if the current action a^1 is repeated T consecutive stages, i.e. $a(t+T) = ... = a(t+2T-1) = a^1$, then $BR_i(\tilde{z}_{-i}(t+2T)) = BR_i(a^1_{-i})$

[6] Since no player is indifferent between distinct strategies, the best response to the current action profile, $BR_i(a^0_{-i})$, is a singleton.

for all players. The probability of such an event is at least $(1-\bar{\varepsilon})^{n(T-1)}$. If the joint action a^1 is an equilibrium, then by Proposition 3.3.1, we are done. Otherwise, there must be at least one player $\mathcal{P}_{i(2)} \in \mathcal{P}$ such that $a^1_{i(2)} \notin BR_{i(2)}(a^1_{-i(2)})$ and hence $a^1_{i(2)} \notin BR_{i(2)}(\tilde{z}_{-i(2)}(t+2T))$.

One can repeat the arguments above to construct a sequence of profiles $a^0, a^1, a^2, ..., a^m$, where $a^k = (a^*_{i(k)}, a^{k-1}_{-i(k)})$ for all $k \geq 1$, with the property that

$$\phi(a^0) < \phi(a^1) < ... < \phi(a^m),$$

and a^m is an equilibrium. This means that given $\{\tilde{z}_{-i}(t)\}_{i=1}^n$, there exist constants

$$\begin{array}{rcl} \tilde{T} & = & (|\mathcal{A}|+1)T > 0, \\ \tilde{\varepsilon} & = & \big(\underline{\varepsilon}(1-\bar{\varepsilon})^{n-1}\big)^{|\mathcal{A}|}\big((1-\bar{\varepsilon})^{n(T-1)}\big)^{|\mathcal{A}|+1} > 0, \end{array}$$

both of which are independent of t, such that the following event happens with probability at least $\tilde{\varepsilon}$: $a(t+\tilde{T})$ is an equilibrium and $a_i(t+\tilde{T}) \in BR_i(\tilde{z}_{-i}(t+\tilde{T}))$ for all players $\mathcal{P}_i \in \mathcal{P}$. This implies that $a(t)$ converges to a pure equilibrium almost surely. □

3.4 Congestion Games and Distributed Traffic Routing

In this section, we illustrate the main results on congestion games, as defined in Section 2.3.3, which are a special case of the generalized ordinal potential games addressed in Theorems 3.2.1 and 3.3.1. We illustrate these results on a simulation of distributed traffic routing. We go on to discuss how to modify player utilities in distributed traffic routing to allow a centralized planner to achieve a desired collective objective through distributed learning.

3.4.1 Distributed Traffic Routing

We consider a congestion game, as defined in Section 2.3.3, with 100 players, or drivers, seeking to traverse from node A to node B along 10 different parallel roads as illustrated in Figure 3.1. Each driver can select any road as a possible route. In terms of congestion games, the set of resources is the set of roads, $\mathcal{R}$, and each player can select one road, i.e., $\mathcal{A}_i = \mathcal{R}$.

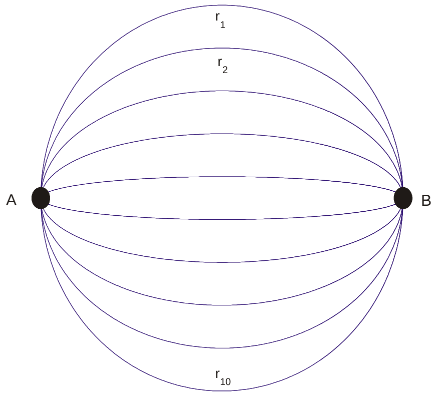


Figure 3.1: Fading Memory JSFP with Inertia: Congestion Game Example – Network Topology

Each road has a quadratic cost function with positive (randomly chosen) coefficients,

$$c_{r_i}(k) = a_i k^2 + b_i k + c_i,\ i = 1, ..., 10,$$

where k represent the number of vehicles on that particular road. The actual coefficients are unimportant as we are just using this example as an opportunity to illustrate the convergence properties of the algorithm fading memory JSFP with inertia. This cost function may represent the delay incurred by a driver as a function of the number of other drivers sharing the same road.

We simulated a case where drivers choose their initial routes randomly, and every day thereafter, adjusted their routes using fading memory JSFP with inertia. The parameters $\alpha_i(t)$ are chosen as 0.5 for all days and all players, and the fading memory parameter ρ is chosen as 0.03. The number of vehicles on each road fluctuates initially and then stabilizes as illustrated in Figure 3.2. Figure 3.3 illustrates the evolution of the congestion cost on each road. One can observe that the congestion cost on each road converges approximately to the same value, which is consistent with a Nash equilibrium with large number of drivers. This behavior resembles an approximate "Wardrop equilibrium" [War52], which represents a steady-state situation in which the congestion cost on each road is equal due to the fact that, as the number of drivers increases, the effect of an individual driver on the traffic conditions becomes negligible.

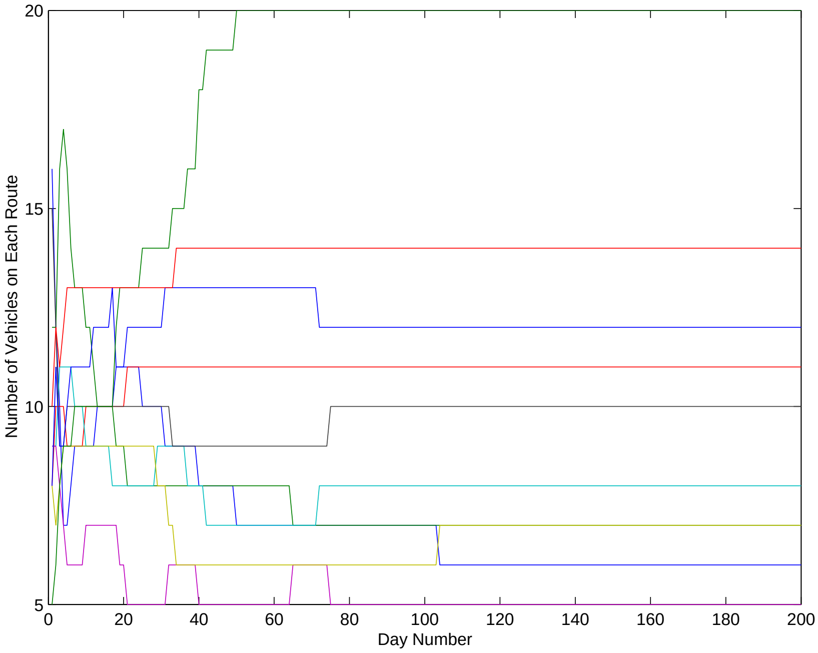


Figure 3.2: Fading Memory JSFP with Inertia: Evolution of Number of Vehicles on Each Route

Note that FP could not be implemented even on this very simple congestion game. A driver using FP would need to track the empirical frequencies of the choices of the 99 other drivers and compute an expected utility evaluated over a probability space of

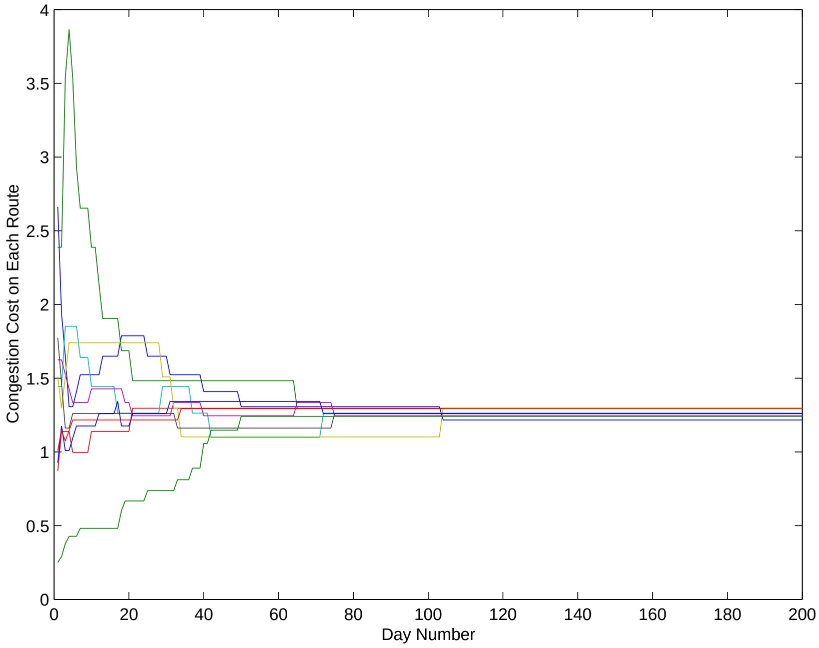


Figure 3.3: Fading Memory JSFP with Inertia: Evolution of Congestion Cost on Each Route

dimension 10^{99}.

It turns out that JSFP, fading memory JSFP, or other virtual payoff based learning algorithms are strongly connected to actual driver behavioral models. Consider the driver adjustment process considered in [BPK91] which is illustrated in Figure 3.4. The adjustment process highlighted is precisely JSFP with Inertia.

3.4.2 Incorporating Tolls to Minimize the Total Congestion

It is well known that a Nash equilibrium may not minimize the total congestion experienced by all drivers [Rou03]. In this section, we show how a global planner can minimize the total congestion by implementing tolls on the network. The results are applicable to general congestion games, but we present the approach in the language of distributed traffic routing.

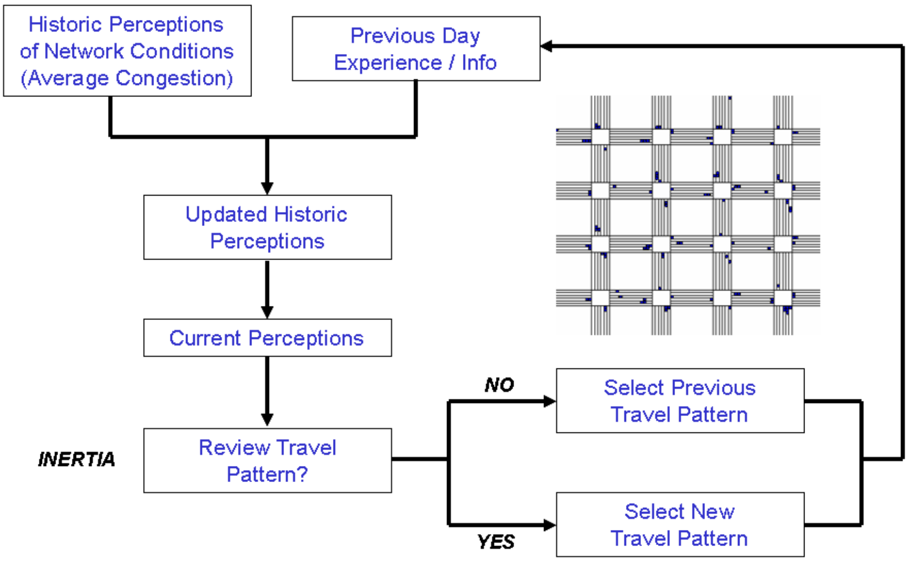

Historic Perceptions of Network Conditions (Average Congestion)
Previous Day Experience / Info
Updated Historic Perceptions
Current Perceptions
INERTIA
Review Travel Pattern?
NO
YES
Select Previous Travel Pattern
Select New Travel Pattern

Figure 3.4: Example of a Driver Adjustment Process

The total congestion experienced by all drivers on the network is

$$T_c(a) := \sum_{r \in \mathcal{R}} \sigma_r(a) c_r(\sigma_r(a)).$$

Define a new congestion game where each driver's utility takes the form

$$U_i(a) = -\sum_{r \in a_i} \big(c_r(\sigma_r(a)) + t_r(\sigma_r(a))\big),$$

where $t_r(\cdot)$ is the toll imposed on road r which is a function of the number of users of road r.

The following proposition, which is a special case of Proposition 5.3.1, outlines how to incorporate tolls so that the minimum congestion solution is a Nash equilibrium. The approach is similar to the taxation approaches for nonatomic congestion games proposed in [Mil04, San02].

Proposition 3.4.1. *Consider a congestion game of any network topology. If the imposed tolls are set as*

$$t_r(k) = (k-1)[c_r(k) - c_r(k-1)], \ \ \forall k \geq 1,$$

then the total negative congestion experienced by all drivers, $\phi_c(a) := -T_c(a)$, is a potential function for the congestion game with tolls.

By implementing the tolling scheme set forth in Proposition 3.4.1, we guarantee that all action profiles that minimize the total congestion experienced on the network are equilibria of the congestion game with tolls. However, there may be addition equilibria at which an inefficient operating condition can still occur. The following proposition establishes the uniqueness of a strict Nash equilibrium for congestion games of parallel network topologies such as the one considered in this example.

Proposition 3.4.2. *Consider a congestion game with nondecreasing congestion functions where each driver is allowed to select any one road, i.e. $\mathcal{A}_i = \mathcal{R}$ for all drivers. If the congestion game has at least one strict equilibrium, then all equilibria have the same aggregate vehicle distribution over the network. Furthermore, all equilibria are strict.*

Proof. Suppose action profiles a^1 and a^2 are equilibria with a^1 being a strict equilibrium. We will use the shorthand notation $\sigma_r^{a^1}$ to represent $\sigma_r(a^1)$. Let $\sigma(a^1) := (\sigma_{r_1}^{a^1}, ..., \sigma_{r_n}^{a^1})$ and $\sigma(a^2) := (\sigma_{r_1}^{a^2}, ..., \sigma_{r_n}^{a^2})$ be the aggregate vehicle distribution over the network for equilibrium a^1 and a^2. If $\sigma(a^1) \neq \sigma(a^2)$, there exists a road a such that $\sigma_a^{a^1} > \sigma_a^{a^2}$ and a road b such that $\sigma_b^{a^1} < \sigma_b^{a^2}$. Therefore, we know that

$$\begin{aligned} c_a(\sigma_a^{a^1}) &\geq c_a(\sigma_a^{a^2} + 1), \\ c_b(\sigma_b^{a^2}) &\geq c_b(\sigma_b^{a^1} + 1). \end{aligned}$$

Since a^1 and a^2 are equilibrium with a^1 being strict,

$$\begin{aligned} c_a(\sigma_a^{a^1}) &< c_{r_i}(\sigma_{r_i}^{a^1}+1), \quad \forall r_i \in \mathcal{R}, \\ c_b(\sigma_b^{a^2}) &\leq c_{r_i}(\sigma_{r_i}^{a^2}+1), \quad \forall r_i \in \mathcal{R}. \end{aligned}$$

Using the above inequalities, we can show that

$$c_a(\sigma_a^{a^1}) \geq c_a(\sigma_a^{a^2}+1) \geq c_b(\sigma_b^{a^2}) \geq c_b(\sigma_b^{a^1}+1) > c_a(\sigma_a^{a^1}),$$

which gives us a contradiction. Therefore $\sigma(a^1) = \sigma(a^2)$. Since a^1 is a strict equilibrium, then a^2 must be a strict equilibrium as well. □

When the tolling scheme set forth in Proposition 3.4.1 is applied to the congestion game example considered previously, the resulting congestion game with tolls is a potential game in which no player is indifferent between distinct strategies. Proposition 3.4.1 guarantees us that the action profiles that minimize the total congestion experienced by all drivers on the network are in fact strict equilibria of the congestion game with tolls. Furthermore, if the new congestion functions are nondecreasing[7], then by Proposition 3.4.2, all strict equilibria must have the same aggregate vehicle distribution over the network, and therefore must minimize the total congestion experienced by all drivers on the network. Therefore, the action profiles generated by fading memory JSFP with inertia converge to an equilibrium that minimizes the total congestion experienced by all users, as shown in Figure 3.5.

3.5 Concluding Remarks and Future Work

We have analyzed the long-term behavior of a large number of players in large-scale games where players are limited in both their observational and computational capabilities. In particular, we analyzed a version of JSFP and showed that it accommodates

[7] Simple conditions on the original congestion functions can be established to guarantee that the new congestion functions, i.e congestion plus tolls, are nondecreasing.

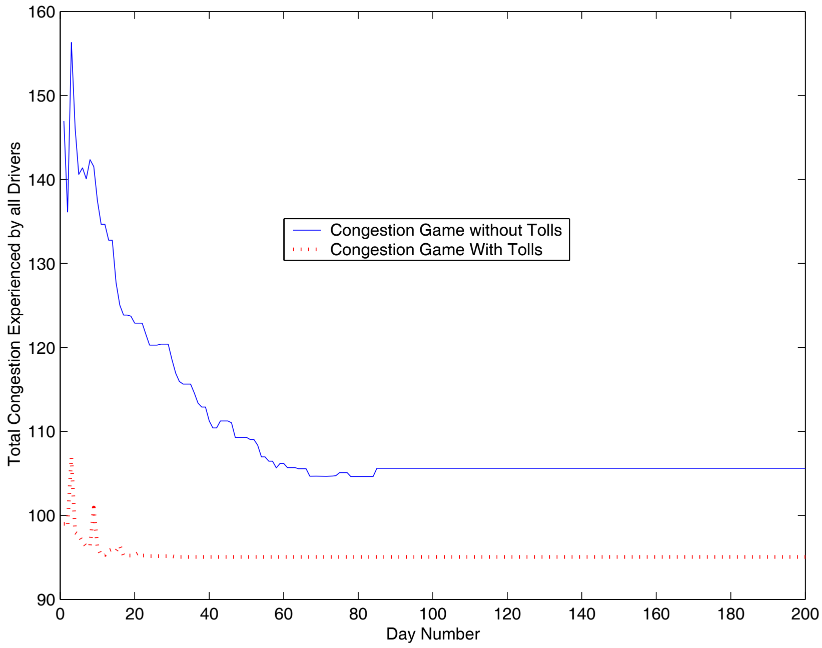


Figure 3.5: Fading Memory JSFP with Inertia: Evolution of Total Congestion Experienced by All Drivers with and without Tolls.

inherent player limitations in information gathering and processing. Furthermore, we showed that JSFP has guaranteed convergence to a pure Nash equilibrium in all generalized ordinal potential games, which includes but is not limited to all congestion games, when players use some inertia either with or without exponential discounting of the historical data. The methods were illustrated on a transportation congestion game, in which a large number of vehicles make daily routing decisions to optimize their own objectives in response to the aggregate congestion on each road of interest. An interesting continuation of this research would be the case where players observe only the actual utilities they receive. This situation will be the focus of Chapter 5.

The method of proof of Theorems 3.2.1 and 3.3.1 relies on inertia to derive a positive probability of a single player seeking to make an utility improvement, thereby increasing the potential function. This suggests a convergence rate that is exponential in the game size, i.e., number of players and actions. It should be noted that inertia is simply a proof device that assures convergence for generic potential games. The proof provides just one out of multiple paths to convergence. The simulations reflect that convergence can be much faster. Indeed, simulations suggest that convergence is possible even in the absence of inertia but not necessarily for all potential games. Furthermore, recent work [HM06] suggests that convergence rates of a broad class of distributed learning processes can be exponential in the game size as well, and so this seems to be a limitation in the framework of distributed learning rather than any specific learning process (as opposed to centralized algorithms for computing an equilibrium).

3.6 Appendix to Chapter 3

3.6.1 Proof of Theorem 3.2.1

This section is devoted to the proof of Theorem 3.2.1. It will be helpful to note the following simple observations:

1. The expression for $U_i(\bar{a}_i, z_{-i}(t))$ in equation (3.4) is linear in $z_{-i}(t)$.
2. If an action profile, $a^0 \in \mathcal{A}$, is repeated over the interval $[t, t+N-1]$, i.e.,

   $$a(t) = a(t+1) = ... = a(t+N-1) = a^0,$$

   then $z(t+N)$ can be written as

   $$z(t+N) = \frac{t}{t+N} z(t) + \frac{N}{t+N} \mathbf{v}^{a^0},$$

   and likewise $z_{-i}(t+N)$ can be written as

   $$z_{-i}(t+N) = \frac{t}{t+N} z_{-i}(t) + \frac{N}{t+N} \mathbf{v}^{a^0_{-i}}.$$

We begin by defining the quantities $\delta_i(t)$, M_u, m_u, and γ as follows. Assume that player $\mathcal{P}_i$ played a best response at least one time in the period $[0, t]$, where $t \in [0, \infty)$. Define

$$\delta_i(t) := \min\{0 \leq \tau \leq t : a_i(t - \tau) \in BR_i(z_i(t - \tau))\}.$$

In other words, $t - \delta_i(t)$ is the last time in the period $[0, t]$ at which player $\mathcal{P}_i$ played a best response. If player $\mathcal{P}_i$ never played a best response in the period $[0, t]$, then we adopt the convention $\delta_i(t) = \infty$. Note that

$$a_i(t - \tau) = a_i(t), \ \forall \tau \in \{0, 1, ..., \min\{\delta_i(t), t\}\}.$$

Now define

$$\begin{aligned}
M_u &:= \max\{|U_i(a^1) - U_i(a^2)| : a^1, a^2 \in \mathcal{A}, \mathcal{P}_i \in \mathcal{P}\}, \\
m_u &:= \min\{|U_i(a^1) - U_i(a^2)| : |U_i(a^1) - U_i(a^2)| > 0, a^1, a^2 \in \mathcal{A}, \mathcal{P}_i \in \mathcal{P}\}, \\
\gamma &:= \lceil M_u / m_u \rceil,
\end{aligned}$$

where $\lceil \cdot \rceil$ denotes integer ceiling.

The proof of fading memory JSFP with inertia relied on a notion of memory dominance. This means that if the current action profile is repeated a sufficient number of times (finite and independent of time) then a best response to the weighted empirical frequencies is equivalent to a best response to the current action profile and hence will increase the potential provided that there is only a unique deviator. This will always happen with at least a fixed (time independent) probability because of the players' inertia.

In the non-discounted case the memory dominance approach will not work for the reason that the probability of dominating the memory because of the players' inertia diminishes with time. However, the following claims show that one does not need to dominate the entire memory, but rather just the portion of time for which the player was playing a suboptimal action. By dominating this portion of the memory, one can guarantee that a unilateral best response to the empirical frequencies will increase the potential. This is the fundamental idea in the proof of Theorem 3.2.1.

Claim 3.6.1. *Consider a player $\mathcal{P}_i$ with $\delta_i(t) < \infty$. Let t_1 be any finite integer satisfying*

$$t_1 \geq \gamma \delta_i(t).$$

If an action profile, $a^0 \in \mathcal{A}$, is repeated over the interval $[t, t+t_1]$, i.e.,

$$a(t) = a(t+1) = \cdots = a(t+t_1) = a^0,$$

then

$$\hat{a}_i \in BR_i(z_{-i}(t+t_1+1)) \Rightarrow U_i(\hat{a}_i, a^0_{-i}) \geq U_i(a^0_i, a^0_{-i}),$$

i.e., player $\mathcal{P}_i$'s best response at time $t+t_1+1$ cannot be a worse response to a^0_{-i} than a^0_i.

Proof. Since $\hat{a}_i \in BR_i(z_{-i}(t+t_1+1))$,

$$U_i(\hat{a}_i, z_{-i}(t+t_1+1)) - U_i(a^0_i, z_{-i}(t+t_1+1)) \geq 0.$$

Expressing $z_{-i}(t+t_1+1)$ as a summation over the intervals $[0, t-\delta_i(t)-1]$, $[t-\delta_i(t), t-1]$, and $[t, t+t_1]$ and using the definition (3.4) leads to

$$\begin{aligned}
&(t_1+1)\big[U_i(\hat{a}_i, a^0_{-i}) - U_i(a^0_i, a^0_{-i})\big] \\
&+ \sum_{\tau=t-\delta_i(t)}^{t-1} \big[U_i(\hat{a}_i, a_{-i}(\tau)) - U_i(a^0_i, a_{-i}(\tau))\big] \\
&+(t-\delta_i(t))\big[U_i(\hat{a}_i, z_{-i}(t-\delta_i(t))) - U_i(a^0_i, z_{-i}(t-\delta_i(t)))\big] \geq 0.
\end{aligned}$$

Now, since

$$a_i(t-\delta_i(t)) = a_i(t-\delta_i(t)+1) = \cdots = a_i(t) = a^0_i \in BR_i(z_{-i}(t-\delta_i(t))),$$

meaning that the third term above is negative, and so

$$\begin{aligned}&(t_1+1)\big[U_i(\hat{a}_i,a^0_{-i})-U_i(a^0_i,a^0_{-i})\big]\\&+\sum_{\tau=t-\delta_i(t)}^{t-1}\big[U_i(\hat{a}_i,a_{-i}(\tau))-U_i(a^0_i,a_{-i}(\tau))\big]\geq 0.\end{aligned}$$

This implies that

$$\big[U_i(\hat{a}_i,a^0_{-i})-U_i(a^0_i,a^0_{-i})\big]\geq -\frac{M_u\delta_i(t)}{t_1+1}>-m_u,$$

or, alternatively,

$$\big[U_i(a^0_i,a^0_{-i})-U_i(\hat{a}_i,a^0_{-i})\big]<m_u.$$

If the quantity in brackets were positive, this would violate the definition of m_u — unless $\hat{a}_i=a^0_i$. In either case,

$$U_i(\hat{a}_i,a^0_{-i})-U_i(a^0_i,a^0_{-i})\geq 0.$$

□

There are certain action profile/empirical frequency values where the next play is "forced". Define the time-dependent (forced-move) set $\mathcal{F}(t)\subset\mathcal{A}\times\Delta(\mathcal{A})$ as

$$(\bar{a},\bar{z})\in\mathcal{F}(t)$$

$$\Leftrightarrow$$

$$\bar{a}_i\in BR_i\left(\frac{t}{t+1}\bar{z}_{-i}+\frac{1}{t+1}\mathbf{v}^{\bar{a}_{-i}}\right),\quad\forall i\in\{1,...,n\}.$$

So the condition $(a(t),z(t))\in\mathcal{F}(t)$, implies that for all i, "today's" action necessarily lies in "tomorrow's" best response, i.e.,

$$a_i(t)\in BR_i(z_{-i}(t+1)).$$

By the rule JSFP-1, the next play $a_i(t+1)=a_i(t)$ is *forced* for all $i\in\{1,...,N\}$.

Now define

$$\pi(t; a(t), z(t)) := \min\left\{\tau \geq 0 : (a(t+\tau), z(t+\tau)) \notin \mathcal{F}(t+\tau)\right\}. \tag{3.6}$$

If this is never satisfied, then set $\pi(t; a(t), z(t)) = \infty$.

For the sake of notational simplicity, we will drop the explicit dependence on $a(t)$ and $z(t)$ and simply write $\pi(t)$ instead of $\pi(t; a(t), z(t))$.

A consequence of the definition of $\pi(t)$ is that for a given $a(t)$ and $z(t)$, 1) $a(t)$ *must* be repeated over the interval $[t, t + \pi(t)]$. Furthermore, at time $t + \pi(t) + 1$, *at least one* player can improve (over yet another repeated play of $a(t)$) by playing a best response at time $t + \pi(t) + 1$. Furthermore, the probability that *exactly one* player will switch to a best response action at time $t + \pi(t) + 1$ is at least $\underline{\varepsilon}(1 - \overline{\varepsilon})^{n-1}$.

The following claim shows that this improvement opportunity remains even if $a(t)$ is repeated for *longer* than $\pi(t)$ (because of inertia).

Claim 3.6.2. *Let $a(t)$ and $z(t)$ be such that $\pi(t) < \infty$. Let t_1 be any integer satisfying $\pi(t) \leq t_1 < \infty$. If*

$$a(t) = a(t+1) = \cdots = a(t+\pi(t)) = \cdots = a(t+t_1),$$

then

$$a_i(t) \notin BR_i(z_{-i}(t+t_1+1)), \text{ for some } i \in \{1, ..., n\}.$$

Proof. Let $i \in \{1, ..., n\}$ be such that

$$a_i(t) \notin BR_i(z_{-i}(t+\pi(t)+1))$$

and

$$a_i(t) \in BR_i(z_{-i}(t+\pi(t))).$$

The existence of such an i is assured by the definition of $\pi(t)$. Pick $\hat{a}_i \in BR_i(z_{-i}(t + \pi(t) + 1))$. We have

$$\begin{aligned}
&U_i(\hat{a}_i, z_{-i}(t + \pi(t) + 1)) - U_i(a_i(t), z_{-i}(t + \pi(t) + 1)) \\
&\quad = [U_i(\hat{a}_i, z_{-i}(t + \pi(t))) - U_i(a_i(t), z_{-i}(t + \pi(t)))] \frac{t + \pi(t)}{t + \pi(t) + 1} \\
&\qquad + [U_i(\hat{a}_i, a_{-i}(t)) - U_i(a_i(t), a_{-i}(t))] \frac{1}{t + \pi(t) + 1} > 0.
\end{aligned}$$

Since $a_i(t) \in BR_i(z_{-i}(t + \pi(t)))$, we must have

$$U_i(\hat{a}_i, a_{-i}(t)) - U_i(a_i(t), a_{-i}(t)) > 0.$$

This implies

$$\begin{aligned}
&U_i(\hat{a}_i, z_{-i}(t + t_1 + 1)) - U_i(a_i(t), z_{-i}(t + t_1 + 1)) \\
&\quad = [U_i(\hat{a}_i, z_{-i}(t + \pi(t) + 1)) - U_i(a_i(t), z_{-i}(t + \pi(t) + 1))] \frac{t + \pi(t) + 1}{t + t_1 + 1} \\
&\qquad + [U_i(\hat{a}_i, a_{-i}(t)) - U_i(a_i(t), a_{-i}(t))] \frac{t_1 - \pi(t)}{t + t_1 + 1} > 0.
\end{aligned}$$

□

Claim 3.6.3. *If, at any time, $a(t)$ is not an equilibrium, then $\pi(t) \leq \gamma t$.*

Proof. Let $a^0 := a(t)$. Since a^0 is not an equilibrium,

$$a_i^0 \notin BR_i(a_{-i}^0), \text{ for some } i \in \{1, ..., n\}.$$

Pick $\hat{a}_i \in BR_i(a_{-i}^0)$ so that $U_i(\hat{a}_i, a_{-i}^0) - U_i(a_i^0, a_{-i}^0) > m_u$. If

$$a(t) = a(t + 1) = \cdots = a(t + \gamma t) = a^0,$$

then

$$\begin{aligned}
&U_i(\hat{a}_i, z_{-i}(t + \gamma t + 1)) - U_i(a_i^0, z_{-i}(t + \gamma t + 1)) \\
&= \frac{t[U_i(\hat{a}_i, z_{-i}(t)) - U_i(a_i^0, z_{-i}(t))] + (\gamma t + 1)[U_i(\hat{a}_i, a_{-i}^0) - U_i(a_i^0, a_{-i}^0)]}{t + \gamma t + 1} \\
&\geq \frac{-tM_u + (\gamma t + 1)m_u}{t + \gamma t + 1} \\
&> 0.
\end{aligned}$$

□

Claim 3.6.4. *Consider a finite generalized ordinal potential game with a potential function $\phi(\cdot)$ with player utilities satisfying Assumption 3.2.2. For any time $t \geq 0$, suppose that*

1. *$a(t)$ is not an equilibrium; and*
2. *$\max_{1 \leq i \leq n} \delta_i(t) \leq \bar{\delta}$ for some $\bar{\delta} \leq t$.*

Define

$$\psi(t) := 1 + \max\left\{\pi(t), \gamma\bar{\delta}\right\}.$$

Then $\psi(t) \leq 1 + \gamma t$ and

$$\mathbf{Pr}\left[\phi(a(t+\psi(t))) > \phi(a(t)) \mid a(t), z(t)\right] \geq \underline{\varepsilon}(1-\overline{\varepsilon})^{n(1+\gamma\bar{\delta})-1},$$

and

$$\max_{1 \leq i \leq n} \delta_i(t+\psi(t)) \leq 1 + (1+\gamma)\bar{\delta}.$$

Proof. Since $a(t)$ is not an equilibrium, Claim 3.6.3 implies that $\pi(t) \leq \gamma t$, which in turn implies the above upper bound on $\psi(t)$.

First consider the case where $\pi(t) \geq \gamma\bar{\delta}$, i.e., $\psi(t) = 1 + \pi(t)$. According to the definition of $\pi(t)$ in equation (3.6), $a(t)$ *must* be repeated as a best response in the period $[t, t+\pi(t)]$. Furthermore, we must have

$$\max_{1 \leq i \leq n} \delta_i(t+\psi(t)) \leq 1$$

and $a_i(t) \notin BR_i(z_{-i}(t+\psi(t)))$ for at least one player $\mathcal{P}_i$. The probability that exactly one such player $\mathcal{P}_i$ will switch to a choice different than $a_i(t)$ at time $t+\psi(t)$ is at least $\underline{\varepsilon}(1-\overline{\varepsilon})^{n-1}$. But, by Claim 3.6.1 and no-indifference Assumption 3.2.2, such an event would cause

$$U_i(a(t+\pi(t)+1)) > U_i(a(t)) \;\Rightarrow\; \phi(a(t+\pi(t)+1)) > \phi(a(t)).$$

Now consider the case where $\pi(t) < \gamma\bar{\delta}$, i.e., $\psi(t) = 1 + \gamma\bar{\delta}$. In this case,

$$\max_{1 \le i \le n} \delta_i(t + \psi(t)) \le 1 + \gamma\bar{\delta} + \bar{\delta}.$$

Moreover, the event

$$a(t) = \cdots = a(t + \gamma\bar{\delta})$$

will occur with probability at least[8] $(1 - \bar{\varepsilon})^{n\gamma\bar{\delta}}$. Conditioned on this event, Claim 3.6.2 provides that exactly one player $\mathcal{P}_i$ will switch to a choice different than $a_i(t)$ at time $t + \psi(t)$ with probability at least $\underline{\varepsilon}(1 - \bar{\varepsilon})^{n-1}$. By Claim 3.6.1 and no-indifference Assumption 3.5, this would cause

$$U_i(a(t + \psi(t))) > U_i(a(t)) \;\Rightarrow\; \phi(a(t + \psi(t))) > \phi(a(t)).$$

□

Proof of Theorem 3.2.1

It suffices to show that there exists a non-zero probability, $\varepsilon^* > 0$, such that the following statement holds. For any $t \ge 0$, $a(t) \in \mathcal{A}$, and $z(t) \in \Delta(\mathcal{A})$, there exists a finite time $t^* \ge t$ such that, for some equilibrium a^*,

$$\mathbf{Pr}\left[a(\tau) = a^*, \forall \tau \ge t^* \mid a(t), \{z_{-i}(t)\}_{i=1}^{n}\right] \ge \varepsilon^*. \tag{3.7}$$

In other words, the probability of convergence to an equilibrium by time t^* is at least ε^*. Since ε^* *does not* depend on t, $a(t)$, or $z(t)$, this will imply that the action profile converges to an equilibrium almost surely.

We will construct a series of events that can occur with positive probability to establish the bound in equation (3.7).

[8] In fact, a tighter bound can be derived by exploiting the forced moves for a duration of $\pi(t)$.

Let $t_0 = t+1$. All players will play a best response at time t_0 with probability at least $\underline{\varepsilon}^n$. Therefore, we have

$$\mathbf{Pr}\left[\max_{1\leq i\leq n} \delta_i(t_0) = 0 \mid a(t), \{z_{-i}(t)\}_{i=1}^n\right] \geq \underline{\varepsilon}^n. \tag{3.8}$$

Assume that $a(t_0)$ is not an equilibrium. Otherwise, according to Proposition 3.2.2, $a(\tau) = a(t_0)$ for all $\tau \geq t_0$.

From Claim 3.6.4, define t_1 and δ_1 as

$$\begin{aligned} \delta_1 &:= 1 + (1+\gamma)\delta_0, \\ t_1 &:= t_0 + 1 + \max\{\pi(t_0), \gamma\delta_0\}, \\ &\leq t_0 + 1 + \gamma t_0 = 1 + (1+\gamma)t_0, \end{aligned}$$

where $\delta_0 := 0$. By Claim 3.6.4,

$$\mathbf{Pr}\left[\phi(a(t_1)) > \phi(a(t_0)) \mid a(t_0), \{z_{-i}(t_0)\}_{i=1}^n\right] \geq \underline{\varepsilon}(1-\overline{\varepsilon})^{n(1+\gamma\delta_0)-1}$$

and

$$\max_{1\leq i\leq n} \delta_i(t_1) \leq \delta_1.$$

Similarly, for $k > 0$ we can recursively define

$$\begin{aligned} \delta_k &:= 1 + (1+\gamma)\delta_{k-1}, \\ &= (1+\gamma)^k\delta_0 + \sum_{j=0}^{k-1}(1+\gamma)^j, \\ &= \sum_{j=0}^{k-1}(1+\gamma)^j, \end{aligned}$$

and

$$\begin{aligned} t_k &:= t_{k-1} + 1 + \max\{\pi(t_{k-1}), \gamma\delta_{k-1}\}, \\ &\leq 1 + (1+\gamma)t_{k-1} \\ &\leq (1+\gamma)^k t_0 + \sum_{j=0}^{k-1}(1+\gamma)^j, \end{aligned}$$

where

$$\mathbf{Pr}\left[\phi(a(t_k)) > \phi(a(t_{k-1})) \mid a(t_{k-1}), \{z_{-i}(t_{k-1})\}_{i=1}^n\right] \geq \underline{\varepsilon}(1-\overline{\varepsilon})^{n(1+\gamma\delta_{k-1})-1}$$

and

$$\max_{1\leq i\leq n} \delta_i(t_k) \leq \delta_k,$$

as long as $a(t_{k-1})$ is not an equilibrium.

Therefore, one can construct a sequence of profiles $a(t_0), a(t_1), ..., a(t_k)$ with the property that $\phi(a(t_0)) < \phi(a(t_1)) < ... < \phi(a(t_k))$. Since in a finite generalized ordinal potential game, $\phi(a(t_k))$ cannot increase indefinitely as k increases, we must have

$$\mathbf{Pr}\left[a(t_k) \text{ is an equilibrium for some } t_k \in [t,\infty) \mid a(t), \{z_{-i}(t)\}_{i=1}^n\right] \geq \underline{\varepsilon}^n \prod_{k=0}^{|\mathcal{A}|-1} \underline{\varepsilon}(1-\overline{\varepsilon})^{n(1+\gamma\delta_k)-1},$$

where $\underline{\varepsilon}^n$ comes from (3.8). Finally, from Claim 3.6.1 and Assumption 3.2.2, the above inequality together with

$$\mathbf{Pr}\left[a(t_k) = \cdots = a(t_k+\gamma\delta_k) \mid a(t_k), \{z_{-i}(t_k)\}_{i=1}^n\right] \geq (1-\overline{\varepsilon})^{n\gamma\delta_k} \geq (1-\overline{\varepsilon})^{n\gamma\delta_{|\mathcal{A}|}}$$

implies that for some equilibrium, a^*,

$$\mathbf{Pr}\left[a(\tau) = a^*,\ \forall \tau \geq t^* \mid a(t), \{z_{-i}(t)\}_{i=1}^n\right] \geq \varepsilon^*,$$

where

$$\begin{aligned} t^* &= t_{|\mathcal{A}|} + \gamma\delta_{|\mathcal{A}|} + 1 = (1+\gamma)^{|\mathcal{A}|}t_0 + \sum_{j=0}^{|\mathcal{A}|}(1+\gamma)^j, \\ \varepsilon^* &= \left(\underline{\varepsilon}^n \prod_{k=0}^{|\mathcal{A}|-1} \underline{\varepsilon}(1-\overline{\varepsilon})^{n(1+\gamma\delta_k)-1}\right)\left((1-\overline{\varepsilon})^{n\gamma\delta_{|\mathcal{A}|}}\right). \end{aligned}$$

Since ε^* does not depend on t this concludes the proof. □

CHAPTER 4

Regret Based Dynamics for Weakly Acyclic Games

No-regret algorithms have been proposed to control a wide variety of multi-agent systems. The appeal of no-regret algorithms is that they are easily implementable in large scale multi-agent systems because players make decisions using only retrospective or "regret based" information. Furthermore, there are existing results proving that the collective behavior will asymptotically converge to a set of points of "no-regret" in any game. We illustrate, through a simple example, that no-regret points need not reflect desirable operating conditions for a multi-agent system. Multi-agent systems often exhibit an additional structure (i.e. being "weakly acyclic") that has not been exploited in the context of no-regret algorithms. In this chapter, we introduce a modification of the traditional no-regret algorithms by (i) exponentially discounting the memory and (ii) bringing in a notion of inertia in players' decision process. We show how these modifications can lead to an entire class of regret based algorithms that provide *almost sure* convergence to a pure Nash equilibrium in any weakly acyclic game.

4.1 Introduction

The applicability of regret based algorithms for multi-agent learning has been studied in several papers [Gor05, Bow04, KV05, BP05, GJ03, AMS07]. The appeal of regret based algorithms is two fold. First of all, regret based algorithms are easily implementable in large scale multi-agent systems when compared with other learning algorithms such as fictitious play [MS96a, JGD01]. Secondly, there is a wide range of algorithms, called "no-regret" algorithms, that guarantee that the collective behavior will asymptotically converge to a set of points of no-regret (also referred to as coarse correlated equilibrium) in any game [You05]. A point of no-regret characterizes a situation for which the average utility that a player actually received is as high as the average utility that the player "would have" received had that player used a different fixed strategy at all previous time steps. No-regret algorithms have been proposed in a variety of settings ranging from network routing problems [BEL06] to structured prediction problems [Gor05].

In the more general regret based algorithms, each player makes a decision using *only* information regarding the regret for each of his possible actions. If an algorithm guarantees that a player's maximum regret asymptotically approaches zero then the algorithm is referred to as a no-regret algorithm. The most common no-regret algorithm is regret matching [HM00]. In regret matching, at each time step, each player plays a strategy where the probability of playing an action is proportional to the positive part of his regret for that action. In a multi-agent system, if all players adhere to a no-regret learning algorithm, such as regret matching, then the group behavior will converge asymptotically to a set of points of no-regret [HM00]. Traditionally, a point of no-regret has been viewed as a desirable or efficient operating condition because each player's average utility is as good as the average utility that any other action would have yielded [KV05]. However, a point of no-regret says little about the performance; hence knowing that the collective behavior of a multi-agent system will converge to a set of points of no-regret in general does not guarantee an efficient operation.

There have been attempts to further strengthen the convergence results of no-regret algorithms for special classes of games. For example, in [JGD01], Jafari et al. showed through simulations that no-regret algorithms provide convergence to a Nash equilibrium in dominance solvable, constant-sum, and general sum 2×2 games. In [Bow04], Bowling introduced a gradient based regret algorithm that guarantees that players' strategies converge to a Nash equilibrium in any 2 player 2 action repeated game. In [BEL06], Blum et al. analyzed the convergence of no-regret algorithms in routing games and proved that behavior will approach a Nash equilibrium in various settings. However, the classes of games considered here cannot fully model a wide variety of multi-agent systems.

It turns out that weakly acyclic games, which generalize potential games [MS96b], are closely related to multi-agent systems [MAS07a]. The connection can be seen by recognizing that in any multi-agent system there is a global objective. Each player is assigned a local utility function that is appropriately aligned with the global objective. It is precisely this alignment that connects the realms of multi-agent systems and weakly acyclic games.

An open question is whether no-regret algorithms converge to a Nash equilibrium in n-player weakly acyclic games. In this chapter, we introduce a modification of the traditional no-regret algorithms that (i) exponentially discounts the memory and (ii) brings in a notion of inertia in players' decision process. We show how these modifications can lead to an *entire class* of regret based algorithms that provide almost sure convergence to a *pure* Nash equilibrium in any weakly acyclic game. It is important to note that convergence to a Nash equilibrium also implies convergence to a no-regret point.

In Section 4.2 we discuss the no-regret algorithm, "regret matching," and illustrate the performance issues involved with no-regret points in a simple 3 player identical interest game. In Section 4.3 we introduce a new class of learning dynamics referred to as regret based dynamics with fading memory and inertia. In Section 4.4 we present some simulation results. Section 4.5 presents some concluding remarks.

4.2 Regret Matching

We consider a repeated matrix game with n-player set $\mathcal{P} := \{\mathcal{P}_1, ..., \mathcal{P}_n\}$, a finite action set $\mathcal{A}_i$ for each player $\mathcal{P}_i \in \mathcal{P}$, and a utility function $U_i : \mathcal{A} \to \mathbb{R}$ for each player $\mathcal{P}_i \in \mathcal{P}$, where $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$.

We introduce regret matching, from [HM00], in which players choose their actions based on their *regret* for not choosing particular actions in the past steps.

Define the average regret of player $\mathcal{P}_i$ for an action $a_i \in \mathcal{A}_i$ at time t as

$$R_i^{a_i}(t) := \frac{1}{t} \sum_{\tau=0}^{t-1} \left(U_i(a_i, a_{-i}(\tau)) - U_i(a(\tau)) \right). \tag{4.1}$$

In other words, player $\mathcal{P}_i$'s average regret for $a_i \in \mathcal{A}_i$ would represent the average improvement in his utility if he had chosen $a_i \in \mathcal{A}_i$ in all past steps and all other players' actions had remained unaltered.

Each player $\mathcal{P}_i$ using regret matching computes $R_i^{a_i}(t)$ for every action $a_i \in \mathcal{A}_i$ using the recursion

$$R_i^{a_i}(t) = \frac{t-1}{t} R_i^{a_i}(t-1) + \frac{1}{t} \left(U_i(a_i, a_{-i}(t)) - U_i(a(t)) \right).$$

Note that, at every step $t > 0$, player $\mathcal{P}_i$ updates all entries in his average regret vector $R_i(t) := \left[R_i^{a_i}(t) \right]_{a_i \in \mathcal{A}_i}$. To update his average regret vector at time t, it is sufficient for player $\mathcal{P}_i$ to observe (in addition to the actual utility received at time $t-1$, $U_i(a(t-1))$) his hypothetical utilities $U_i(a_i, a_{-i}(t-1))$, for all $a_i \in \mathcal{A}_i$, that would have been received if he had chosen a_i (instead of $a_i(t-1)$) and all other player actions $a_{-i}(t-1)$ had remained unchanged at step $t-1$.

In regret matching, once player $\mathcal{P}_i$ computes his average regret vector, $R_i(t)$, he chooses an action $a_i(t)$, $t > 0$, according to the probability distribution $p_i(t)$ defined as

$$p_i^{a_i}(t) = \mathbf{Pr}\left[a_i(t) = a_i \right] = \frac{\left[R_i^{a_i}(t) \right]^+}{\sum_{\tilde{a}_i \in \mathcal{A}_i} \left[R_i^{\tilde{a}_i}(t) \right]^+},$$

for any $a_i \in \mathcal{A}_i$, provided that the denominator above is positive; otherwise, $p_i(t)$ is the uniform distribution over $\mathcal{A}_i$ ($p_i(0) \in \Delta(\mathcal{A}_i)$ is always arbitrary). Roughly speaking, a player using regret matching chooses a particular action at any step with probability proportional to the average regret for not choosing that particular action in the past steps. If all players use regret matching, the empirical distribution of the joint actions converge almost surely to the set of coarse correlated equilibria (similar results hold for different regret based adaptive dynamics); see [HM00, HM01, HM03a]. Note that this does not mean that the action profiles $a(t)$ will converge, nor does it mean that the empirical frequencies of $a(t)$ will converge to a point in $\Delta(\mathcal{A})$.

4.2.1 Coarse Correlated Equilibria and No-Regret

The set of coarse correlated equilibrium has a strong connection to the notion of regret.

We will restate the definitions of the joint and marginal empirical frequencies originally defined in Section 3.2. Define the empirical frequency of the joint actions, $z^a(t)$, as the percentage of stages at which all players chose the joint action profile $a \in \mathcal{A}$ up to time $t-1$, i.e.,

$$z^a(t) := \frac{1}{t}\sum_{\tau=0}^{t-1} I\{a(\tau) = a\}.$$

Let $z(t)$ denote the empirical frequency vector formed by the components $\{z^a(t)\}_{a\in\mathcal{A}}$. Note that the dimension of $z(t)$ is the cardinality of the set $\mathcal{A}$, i.e., $|\mathcal{A}|$, and $z(t) \in \Delta(\mathcal{A})$.

Similarly, let $z_{-i}^{a_{-i}}(t)$ be the percentage of stages at which players other then player $\mathcal{P}_i$ have chosen the joint action profile $a_{-i} \in \mathcal{A}_{-i}$ up to time $t-1$, i.e.,

$$z_{-i}^{a_{-i}}(t) := \frac{1}{t}\sum_{\tau=0}^{t-1} I\{a_{-i}(\tau) = a_{-i}\}, \tag{4.2}$$

which, given $z(t)$, can also be expressed as

$$z_{-i}^{a_{-i}}(t) = \sum_{a_i \in \mathcal{A}_i} z^{(a_i, a_{-i})}(t).$$

Let $z_{-i}(t)$ denote the empirical frequency vector formed by the components $\{z_{-i}^{a_{-i}}(t)\}_{a_{-i} \in \mathcal{A}_{-i}}$. Note that the dimension of $z_{-i}(t)$ is the cardinality $|\mathcal{A}_{-i}|$ and $z_{-i}(t) \in \Delta(\mathcal{A}_{-i})$.

Given a joint distribution $z(t)$, the expected utility of player $\mathcal{P}_i$ is

$$\begin{aligned} U_i(z(t)) &= \sum_{a \in \mathcal{A}} U_i(a) z^a(t), \\ &= \frac{1}{t} \sum_{\tau=0}^{t-1} U_i(a(\tau)), \end{aligned}$$

which is precisely the average utility that player $\mathcal{P}_i$ has received up to time $t-1$. The expected utility of player $\mathcal{P}_i$ for any action $a_i \in \mathcal{A}_i$ is

$$\begin{aligned} U_i(a_i, z_{-i}(t)) &= \sum_{a_{-i} \in \mathcal{A}_{-i}} U_i(a_i, a_{-i}) z_{-i}^{a_{-i}}(t), \\ &= \frac{1}{t} \sum_{\tau=0}^{t-1} U_i(a_i, a_{-i}(\tau)), \end{aligned}$$

which is precisely the average utility that player $\mathcal{P}_i$ would have received up to time $t-1$ if player $\mathcal{P}_i$ had played action a_i all previous time periods provided that the other players actions remained unchanged. Therefore, the regret of player $\mathcal{P}_i$ for action $a_i \in \mathcal{A}_i$ at time t can be expressed as

$$R_i^{a_i}(t) = U_i(a_i, z_{-i}(t)) - U_i(z(t)).$$

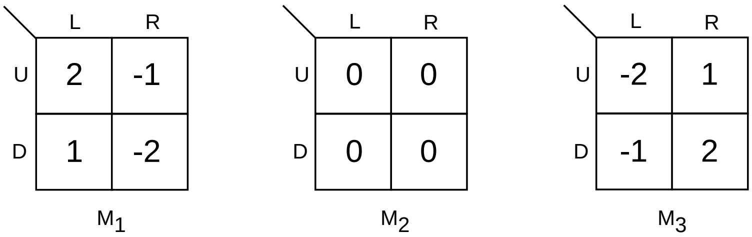
If all players use regret matching, then we know that the empirical frequency $z(t)$ of the joint actions converges almost surely to the set of coarse correlated equilibria. If $z(t)$ is a coarse correlated equilibrium, then we know that for any player $\mathcal{P}_i \in \mathcal{P}$ and any action $a_i \in \mathcal{A}_i$,

$$U_i(a_i, z_{-i}(t)) \leq U_i(z(t)) \;\Rightarrow\; R_i^{a_i}(t) \leq 0.$$

Therefore, stating that the empirical frequency of the joint actions converge to the set of coarse correlated equilibria is equivalent to saying that a player's average regret for any action will asymptotically vanish.

4.2.2 Illustrative Example

In general, the set of Nash equilibria is a proper subset of the set of coarse correlated equilibria. Consider for example the following $3-$player identical interest game characterized by the player utilities shown in Figure 4.1.



M1

	L	R
U	2	-1
D	1	-2

M2

	L	R
U	0	0
D	0	0

M3

	L	R
U	-2	1
D	-1	2

Figure 4.1: A $3-$player Identical Interest Game.

Player $\mathcal{P}_1$ chooses a row U or D, Player $\mathcal{P}_2$ chooses a column L or R, Player $\mathcal{P}_3$ chooses a matrix M_1, or M_2, or M_3. There are two pure Nash equilibria (U, L, M_1) and (D, R, M_3) both of which yield maximum utility 2 to all players. The set of coarse correlated equilibria contains these two pure Nash equilibria as the extremum points of $\Delta(\mathcal{A})$ as well as many other probability distributions in $\Delta(\mathcal{A})$. In particular, the set of coarse correlated equilibria contains the following

$$\left\{ z \in \Delta(\mathcal{A}) : \sum_{a \in \mathcal{A}: a_3 = M_2} z^a = 1,\ z^{ULM_2} = z^{DRM_2},\ z^{URM_2} = z^{DLM_2} \right\}.$$

Any coarse correlated equilibrium of this form yields an expected utility of 0 to all players. Clearly, one of the two pure Nash equilibria would be more desirable to all players then any other outcome including the above coarse correlated equilibria. However, the existing results at the time of writing this dissertation such as Theorem 3.1 in [You05] only guarantee that regret matching will lead players to the set of coarse correlated equilibria and not necessarily to a pure Nash equilibrium. While this example is simplistic in nature, one must believe that situations like this could easily arise in more general weakly acyclic games.

We should emphasize that regret matching could indeed be convergent to a pure Nash equilibrium in weakly acyclic games; however, to the best of authors' knowledge, no proof for such a statement exists. The existing results characterize the long-term behavior of regret matching in general games as convergence to the set of coarse correlated equilibria, whereas we are interested in proving that the action profiles, $a(k)$, generated by regret matching will converge to a pure Nash equilibrium when player utilities constitute a weakly acyclic game, an objective which we will pursue in the next section.

4.3 Regret Based Dynamics with Fading Memory and Inertia

To enable convergence to a pure Nash equilibrium in weakly acyclic games, we will modify the conventional regret based dynamics in two ways. First, we will assume that each player has a fading memory, that is, each player exponentially discounts the influence of its past regret in the computation of its average regret vector. More precisely, each player computes a discounted average regret vector according to the recursion

$$\tilde{R}_i^{\bar{a}_i}(t+1) = (1-\rho)\tilde{R}_i^{\bar{a}_i}(t) + \rho \left(U_i(\bar{a}_i, a_{-i}(t)) - U_i(a(t)) \right),$$

for all $\bar{a}_i \in \mathcal{A}_i$, where $\rho \in (0,1]$ is a parameter with $1-\rho$ being the discount factor, and $\tilde{R}_i^{\bar{a}_i}(1) = 0$.

Second, we will assume that each player chooses an action based on its discounted average regret using some inertia. Therefore, each player $\mathcal{P}_i$ chooses an action $a_i(t)$, at step $t > 1$, according to the probability distribution

$$\alpha_i(t) RB_i(\tilde{R}_i(t)) + (1 - \alpha_i(t))\mathbf{v}^{a_i(t-1)},$$

where $\alpha_i(t)$ is a parameter representing player $\mathcal{P}_i$'s willingness to optimize at time t, $\mathbf{v}^{a_i(t-1)}$ is the vertex of $\Delta(\mathcal{A}_i)$ corresponding to the action $a_i(t-1)$ chosen by player $\mathcal{P}_i$ at step $t-1$, and $RB_i : \mathbb{R}^{|\mathcal{A}_i|} \to \Delta(\mathcal{A}_i)$ is any continuous function (on $\{x \in \mathbb{R}^{|\mathcal{A}_i|} : [x]^+ \neq 0\}$) satisfying

$$\begin{gathered} x^\ell > 0 \Leftrightarrow RB_i^\ell(x) > 0 \\ \text{and} \\ [x]^+ = 0 \Rightarrow RB_i^\ell(x) = \tfrac{1}{|\mathcal{A}_i|},\ \forall \ell, \end{gathered} \tag{4.3}$$

where x^ℓ and $RB_i^\ell(x)$ are the ℓ-th components of x and $RB_i(x)$ respectively.

We will call the above dynamics regret based dynamics (RB) with fading memory and inertia. One particular choice for the function RB_i is

$$RB_i^\ell(x) = \frac{\left[x^\ell\right]^+}{\sum_{m=1}^{|\mathcal{A}_i|} [x^m]^+},\ \text{(when } [x]^+ \neq 0) \tag{4.4}$$

which leads to regret matching with fading memory and inertia. Another particular choice is

$$RB_i^\ell(x) = \frac{e^{\frac{1}{\tau}x^\ell}}{\sum_{x^m>0} e^{\frac{1}{\tau}x^m}} I\{x^\ell > 0\},\ \text{(when } [x]^+ \neq 0),$$

where $\tau > 0$ is a parameter. Note that, for small values of τ, player $\mathcal{P}_i$ would choose, with high probability, the action corresponding to the maximum regret. This choice leads to a stochastic variant of an algorithm called Joint Strategy Fictitious Play with fading memory and inertia; see Section 3.3. Also, note that, for large values of τ, player $\mathcal{P}_i$ would choose any action having positive regret with equal probability.

According to these rules, player $\mathcal{P}_i$ will stay with his previous action $a_i(t-1)$ with probability $1 - \alpha_i(t)$ regardless of his regret. We make the following standing assumption on the players' willingness to optimize.

Assumption 4.3.1. *There exist constants $\underline{\varepsilon}$ and $\bar{\varepsilon}$ such that*

$$0 < \underline{\varepsilon} < \alpha_i(t) < \bar{\varepsilon} < 1$$

for all steps $t > 1$ and for all $i \in \{1, ..., n\}$.

This assumption implies that players are always willing to optimize with some nonzero inertia[1]. A motivation for the use of inertia is to instill a degree of hesitation into the decision making process to ensure that players do not overreact to various situations. We will assume that no player is indifferent between distinct strategies [2].

Assumption 4.3.2. *Player utilities satisfy*

$$U_i(a_i^1, a_{-i}) \neq U_i(a_i^2, a_{-i}), \forall\, a_i^1, a_i^2 \in \mathcal{A}_i, a_i^1 \neq a_i^2,\ \forall\, a_{-i} \in \mathcal{A}_{-i},\ \forall\, i \in \{1, ..., n\}.$$

The following theorem establishes the convergence of regret based dynamics with fading memory and inertia to a pure Nash equilibrium.

Theorem 4.3.1. *In any weakly acyclic game satisfying Assumption 4.3.2, the action profiles $a(t)$ generated by regret based dynamics with fading memory and inertia satisfying Assumption 4.3.1 converge to a pure Nash equilibrium almost surely.*

We provide a complete proof for the above result in the Appendix of this chapter. We note that, in contrast to the existing weak convergence results for regret matching in general games, the above result characterizes the long-term behavior of regret based dynamics with fading memory and inertia, in a strong sense, albeit in a restricted class of games. We next numerically verify our theoretical result through some simulations.

[1] This assumption can be relaxed to holding for sufficiently large t, as opposed to all t.

[2] One could alternatively assume that all pure Nash equilibrium are strict.

4.4 Simulations

4.4.1 Three Player Identical Interest Game

We extensively simulated the RB iterations for the game considered in Figure 4.1. We used the RB_i function given in (4.4) with inertia factor $\alpha = 0.5$ and discount factor $\rho = 0.1$. In all cases, player action profiles $a(t)$ converged to one of the pure Nash equilibria as predicted by our main theoretical result. A typical simulation run shown in Figure 4.2 illustrates the convergence of RB iterations to the pure Nash equilibrium (D, R, M_3).

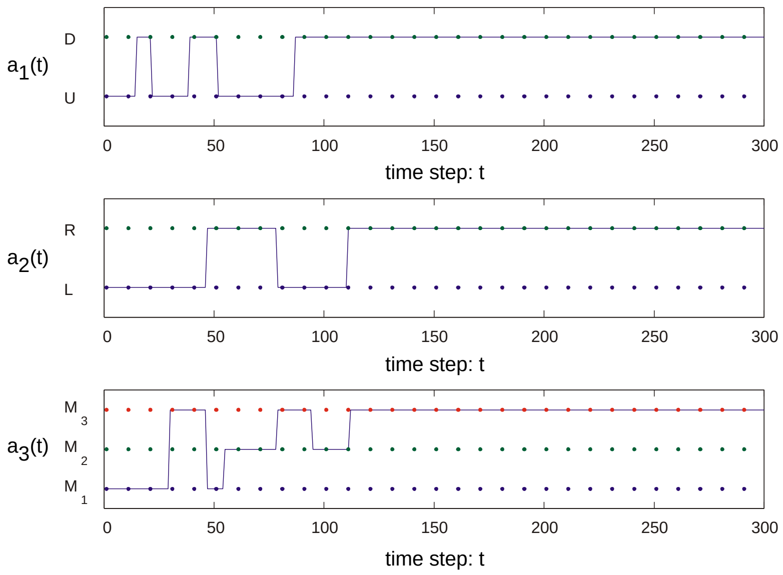


Figure 4.2: Evolution of the actions of players using RB.

4.4.2 Distributed Traffic Routing

We consider a simple congestion game, as defined in Section 2.3.3, with 100 players seeking to traverse from node A to node B along 5 different parallel roads as illustrated in Figure 4.3. Each player can select any road as a possible route. In terms of congestion games, the set of resources is the set of roads, $\mathcal{R}$, and each player can select one road, i.e., $\mathcal{A}_i = \mathcal{R}$.

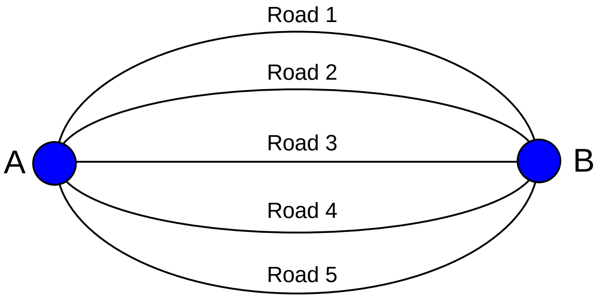


Figure 4.3: Regret Based Dynamics with Inertia: Congestion Game Example – Network Topology

We will assume that each road has a linear cost function with positive (randomly chosen) coefficients,

$$c_{r_i}(k) = a_i k + b_i,\ i = 1, ..., 5,$$

where k represent the number of vehicles on that particular road. This cost function may represent the delay incurred by a driver as a function of the number of other drivers sharing the same road. The actual coefficients or structural form of the cost function are unimportant as we are just using this example as an opportunity to illustrate the convergence properties of the proposed regret based algorithms.

We simulated a case where drivers choose their initial routes randomly, and every day thereafter, adjusted their routes using the regret based dynamics with the RB_i function given in (4.4) with inertia factor $\alpha = 0.85$ and discount factor $\rho = 0.1$. The number of vehicles on each road fluctuates initially and then stabilizes as illustrated in Figure 4.4. Figure 4.5 illustrates the evolution of the congestion cost on each road. One can observe that the congestion cost on each road converges approximately to the same value, which is consistent with a Nash equilibrium with large number of drivers. This behavior resembles an approximate "Wardrop equilibrium" [War52], which represents a steady-state situation in which the congestion cost on each road is equal due to the fact that, as the number of drivers increases, the effect of an individual driver on the traffic conditions becomes negligible.

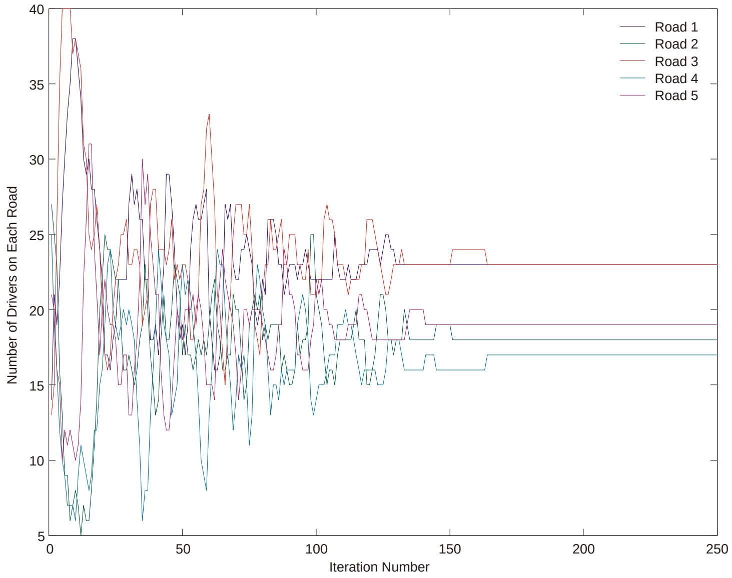


Figure 4.4: Regret Based Dynamics with Inertia: Evolution of Number of Vehicles on Each Route

We would like to note that the simplistic nature of this example was solely for illustrative purposes. Regret based dynamics could be employed on any congestion game with arbitrary network topology and congestion functions. Furthermore, well known learning algorithms such as fictitious play [MS96a] could not be implemented even on this very simple congestion game. A driver using fictitious play would need

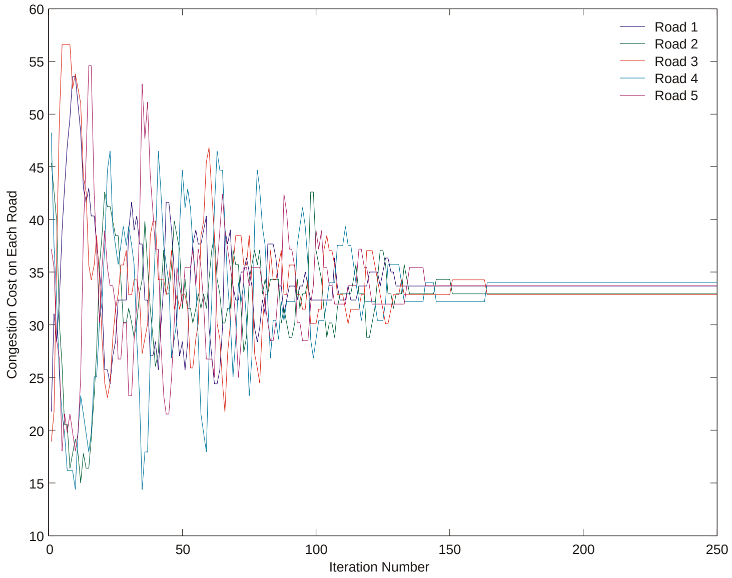


Figure 4.5: Regret Based Dynamics with Inertia: Evolution of Congestion Cost on Each Route

to track the empirical frequencies of the choices of the 99 other drivers and compute an expected utility evaluated over a probability space of dimension 5^{99}.

We would also like to note that in a congestion game, it may be unrealistic to assume that players are aware of the congestion function on each road. This implies that each driver is unaware of his own utility function. However, even in this setting, regret based dynamics can be effectively employed under the condition that each player can evaluate congestion levels on alternative routes. On the other hand, if a player is only aware of the congestion experienced, then one would need to examine the applicability of payoff based algorithms [MYA07] which will be discussed in detail in the following chapter.

4.5 Concluding Remarks and Future Work

In this chapter we analyzed the applicability of regret based algorithms on multi-agent systems. We demonstrated that a point of no-regret may not necessarily be a desirable operating condition. Furthermore, the existing results on regret based algorithms do not preclude these inferior operating points. Therefore, we introduced a modification of the traditional no-regret algorithms that (i) exponentially discounts the memory and (ii) brings in a notion of inertia in players' decision process. We showed how these modifications can lead to an entire class of regret based algorithms that provide convergence to a pure Nash equilibrium in any weakly acyclic game. We believe that similar results hold for no-regret algorithms without fading memory and inertia but thus far the proofs have been elusive.

4.6 Appendix to Chapter 4

4.6.1 Proof of Theorem 4.3.1

We will first state and prove a series of claims. The first claim states that if at any time a player plays an action with positive regret, then the player will play an action with positive regret at all subsequent time steps.

Claim 4.6.1. *Fix any $t_0 > 1$. Then,*

$$\tilde{R}_i^{a_i(t_0)}(t_0) > 0 \Rightarrow \tilde{R}_i^{a_i(t)}(t) > 0$$

for all $t > t_0$.

Proof. Suppose $\tilde{R}_i^{a_i(t_0)}(t_0) > 0$. We have

$$\tilde{R}_i^{a_i(t_0)}(t_0 + 1) = (1 - \rho)\tilde{R}_i^{a_i(t_0)}(t_0) > 0.$$

If $a_i(t_0+1) = a_i(t_0)$, then

$$\tilde{R}_i^{a_i(t_0+1)}(t_0+1) = \tilde{R}_i^{a_i(t_0)}(t_0+1) > 0.$$

If $a_i(t_0+1) \neq a_i(t_0)$, then

$$\tilde{R}_i^{a_i(t_0+1)}(t_0+1) > 0.$$

The argument can be repeated to show that $\tilde{R}_i^{a_i(t)}(t) > 0$, for all $t > t_0$. □

Define

$$\begin{aligned}
M_u &:= \max\{U_i(a) : a \in \mathcal{A}, \mathcal{P}_i \in \mathcal{P}\}, \\
m_u &:= \min\{U_i(a) : a \in \mathcal{A}, \mathcal{P}_i \in \mathcal{P}\}, \\
\delta &:= \min\{|U_i(a^1) - U_i(a^2)| > 0 : \\
&\qquad a^1, a^2 \in \mathcal{A}, a^1_{-i} = a^2_{-i}, \mathcal{P}_i \in \mathcal{P}\}, \\
N &:= \min\{n \in \{1, 2, ...\} : \\
&\qquad (1-(1-\rho)^n)\delta - (1-\rho)^n(M_u - m_u) > \delta/2\}, \\
f &:= \min\{RB_i^m(x) : |x^\ell| \leq M_u - m_u, \forall \ell, \\
&\qquad x^m \geq \delta/2, \text{ for one } m, \ \forall \mathcal{P}_i \in \mathcal{P}\}.
\end{aligned}$$

Note that $\delta, f > 0$, and $|\tilde{R}_i^{a_i}(t)| \leq M_u - m_u$, for all $\mathcal{P}_i \in \mathcal{P}$, $a_i \in \mathcal{A}_i$, $t > 1$.

The second claim states a condition describing the absorptive properties of a strict Nash equilibrium.

Claim 4.6.2. *Fix $t_0 > 1$. Assume*

1. *$a(t_0)$ is a strict Nash equilibrium, and*

2. *$\tilde{R}_i^{a_i(t_0)}(t_0) > 0$ for all $\mathcal{P}_i \in \mathcal{P}$, and*

3. *$a(t_0) = a(t_0+1) = ... = a(t_0+N-1)$.*

Then, $a(t) = a(t_0)$*, for all* $t \geq t_0$*.*

Proof. For any $\mathcal{P}_i \in \mathcal{P}$ and any $a_i \in \mathcal{A}_i$, we have

$$\begin{aligned} \tilde{R}_i^{a_i}(t_0 + N) &= (1-\rho)^N \tilde{R}_i^{a_i}(t_0) \\ &\quad + \big(1 - (1-\rho)^N\big)\big(U_i(a_i, a_{-i}(t_0)) \\ &\quad - U_i(a_i(t_0), a_{-i}(t_0))\big). \end{aligned}$$

Since $a(t_0)$ is a strict Nash equilibrium, for any $\mathcal{P}_i \in \mathcal{P}$ and any $a_i \in \mathcal{A}_i$, $a_i \neq a_i(t_0)$, we have

$$U_i(a_i, a_{-i}(t_0)) - U_i(a_i(t_0), a_{-i}(t_0)) \leq -\delta.$$

Therefore, for any $\mathcal{P}_i \in \mathcal{P}$ and any $a_i \in \mathcal{A}_i$, $a_i \neq a_i(t_0)$,

$$\begin{aligned} \tilde{R}_i^{a_i}(t_0 + N) &\leq (1-\rho)^N (M_u - m_u) - (1 - (1-\rho)^N)\delta \\ &< -\delta/2 < 0. \end{aligned}$$

We also know that, for all $\mathcal{P}_i \in \mathcal{P}$,

$$\tilde{R}_i^{a_i(t_0)}(t_0 + N) = (1-\rho)^N \tilde{R}_i^{a_i(t_0)}(t_0) > 0.$$

This proves the claim. □

The third claim states an event, and associated probability, where the ensuing joint action is a better response to the current joint action profile.

Claim 4.6.3. *Fix* $t_0 > 1$*. Assume*

1. $a(t_0)$ *is not a Nash equilibrium, and*

2. $a(t_0) = a(t_0 + 1) = ... = a(t_0 + N - 1)$

Let $a^* = (a_i^*, a_{-i}(t_0))$ *be such that*

$$U_i(a_i^*, a_{-i}(t_0)) > U_i(a_i(t_0), a_{-i}(t_0)),$$

for some $\mathcal{P}_i \in \mathcal{P}$ *and some* $a_i^* \in \mathcal{A}_i$*. Then,* $\tilde{R}_i^{a_i^*}(t_0 + N) > \delta/2$*, and* a^* *will be chosen at step* $t_0 + N$ *with at least probability* $\gamma := (1 - \overline{\epsilon})^{n-1}\underline{\epsilon}f$*.*

Proof. We have

$$\begin{aligned}\tilde{R}_i^{a_i^*}(t_0 + N) &\geq -(1-\rho)^N(M_u - m_u) + (1 - (1-\rho)^N)\delta \\ &> \delta/2.\end{aligned}$$

Therefore, the probability of player $\mathcal{P}_i$ choosing a_i^* at step $t_0 + N$ is at least $\underline{\epsilon}f$. Because of players' inertia, all other players will repeat their actions at step $t_0 + N$ with probability at least $(1 - \overline{\epsilon})^{n-1}$. This means that the action profile a^* will be chosen at step $t_0 + N$ with probability at least $(1 - \overline{\epsilon})^{n-1}\underline{\epsilon}f$. □

The fourth claim identifies a particular event, and associated probability, guaranteeing that each player will only play actions with positive regret as discussed in Claim 4.6.1.

Claim 4.6.4. *Fix* $t_0 > 1$*. We have* $\tilde{R}_i^{a_i(t)}(t) > 0$ *for all* $t \geq t_0 + 2Nn$ *and for all* $\mathcal{P}_i \in \mathcal{P}$ *with probability at least*

$$\prod_{i=1}^{n} \frac{1}{|\mathcal{A}_i|}\gamma(1 - \overline{\epsilon})^{2Nn}.$$

Proof. Let $a^0 := a(t_0)$. Suppose $\tilde{R}_i^{a_i^0}(t_0) \leq 0$. Furthermore, suppose that a^0 is repeated N consecutive times, i.e. $a(t_0) = ... = a(t_0 + N - 1) = a^0$, which occurs with at least probability at least $(1 - \overline{\epsilon})^{n(N-1)}$.

If there exists a $a^* = (a_i^*, a_{-i}^0)$ such that $U_i(a^*) > U_i(a^0)$, then, by Claim 4.6.3, $\tilde{R}_i^{a_i^*}(t_0 + N) > \delta/2$ and a^* will be chosen at step $t_0 + N$ with at least probability γ. Conditioned on this, we know from Claim 4.6.1 that $\tilde{R}_i^{a_i(t)}(t) > 0$ for all $t \geq t_0 + N$.

If there does not exist such an action a^*, then $\tilde{R}_i^{a_i}(t_0+N) \leq 0$ for all $a_i \in \mathcal{A}_i$. An action profile (a_i^w, a_{-i}^0) with $U_i(a_i^w, a_{-i}^0) < U_i(a^0)$ will be chosen at step t_0+N with at least probability $\frac{1}{|\mathcal{A}_i|}(1-\bar{\epsilon})^{n-1}$. If $a(t_0+N) = (a_i^w, a_{-i}^0)$, and if furthermore (a_i^w, a_{-i}^0) is repeated N consecutive times, i.e., $a(t_0+N) = ... = a(t_0+2N-1)$, which happens with probability at least $(1-\bar{\epsilon})^{n(N-1)}$, then, by Claim 4.6.3, $\tilde{R}_i^{a_i^0}(t_0+2N) > \delta/2$ and the action profile a^0 will be chosen at step (t_0+2N) with at least probability γ. Conditioned on this, we know from Claim 4.6.1 that $\tilde{R}_i^{a_i(t)}(t) > 0$ for all $t \geq t_0+2N$.

In summary, $\tilde{R}_i^{a_i(t)}(t) > 0$ for all $t \geq t_0+2N$ with at least probability

$$\frac{1}{|\mathcal{A}_i|}\gamma(1-\bar{\epsilon})^{2Nn}.$$

We can repeat this argument for each player to show that $\tilde{R}_i^{a_i(t)}(t) > 0$ for all times $t \geq t_0+2Nn$ and for all $\mathcal{P}_i \in \mathcal{P}$ with probability at least

$$\prod_{i=1}^{n}\frac{1}{|\mathcal{A}_i|}\gamma(1-\bar{\epsilon})^{2Nn}.$$

□

FINAL STEP: Establishing convergence to a strict Nash equilibrium:

Proof. Fix $t_0 > 1$. Define $t_1 := t_0+2Nn$. Let $a^1, a^2, \ldots, a^L$ be a finite sequence of action profiles satisfying the conditions given in Subsection 2.3.4 with $a^1 := a(t_1)$.

Suppose $\tilde{R}_i^{a_i(t)}(t) > 0$ for all $t \geq t_1$ and for all $\mathcal{P}_i \in \mathcal{P}$, which, by Claim 4.6.4, occurs with probability at least

$$\prod_{i=1}^{n}\frac{1}{|\mathcal{A}_i|}\gamma(1-\bar{\epsilon})^{2Nn}.$$

Suppose further that $a(t_1) = ... = a(t_1+N-1) = a^1$ which occurs with at least probability $(1-\bar{\epsilon})^{n(N-1)}$. According to Claim 4.6.3 the action profile a^2 will be played at step $t_2 := t_1+N$ with at least probability γ. Suppose now $a(t_2) = ... = a(t_2+$

$N-1) = a^2$, which occurs with at least probability $(1-\overline{\epsilon})^{n(N-1)}$. According to Claim 4.6.3, the action profile a^3 will be played at step $t_3 := t_2 + N$ with at least probability γ.

We can repeat the above arguments until we reach the strict Nash equilibrium a^L at step t_L (recursively defined as above) and stay at a^L for N consecutive steps. From Claim 2, this would mean that the action profile would stay at a^L for all $t \geq t_L$.

Therefore, given $t_0 > 1$, there exists constants $\tilde{\epsilon} > 0$ and $\tilde{T} > 0$, both of which are independent of t_0, and a strict Nash equilibrium a^*, such that the following event happens with at least probability $\tilde{\epsilon}$: $a(t) = a^*$ for all $t \geq t_0 + \tilde{T}$. This proves Theorem 4.1.

□

CHAPTER 5

Payoff Based Dynamics for Weakly Acyclic Games

We consider repeated multi-player games in which players repeatedly and simultaneously choose strategies from a finite set of available strategies according to some strategy adjustment process. We focus on the specific class of weakly acyclic games, which is particularly relevant for multi-agent cooperative control problems. A strategy adjustment process determines how players select their strategies at any stage as a function of the information gathered over previous stages. Of particular interest are “payoff based” processes, in which at any stage, players only know their own actions and (noise corrupted) payoffs from previous stages. In particular, players do not know the actions taken by other players and do not know the structural form of payoff functions. We introduce three different payoff based processes for increasingly general scenarios and prove that after a sufficiently large number of stages, player actions constitute a Nash equilibrium at any stage with arbitrarily high probability. We also show how to modify player utility functions through tolls and incentives in so-called congestion games, a special class of weakly acyclic games, to guarantee that a centralized objective can be realized as a Nash equilibrium. We illustrate the methods with a simulation of distributed routing over a network.

5.1 Introduction

The objective in distributed cooperative control for multi-agent systems is to enable a collection of “self-interested” agents to achieve a desirable “collective” objective. There are two overriding challenges to achieving this objective. The first is complexity: finding an optimal solution by a centralized algorithm may be prohibitively difficult when there are large numbers of interacting agents. This motivates the use of adaptive methods that enable agents to “self organize” into suitable, if not optimal, collective solutions.

The second challenge is limited information. Agents may have limited knowledge about the status of other agents, except perhaps for a small subset of “neighboring” agents. An example is collective motion control for mobile sensor platforms (e.g., [GSM05]). In these problems, mobile sensors seek to position themselves to achieve various collective objectives such as rendezvous or area coverage. Sensors can communicate with neighboring sensors, but otherwise do not have global knowledge of the domain of operation or the status and locations of non-neighboring sensors.

A typical assumption is that agents are endowed with a reward or utility function that depends on their own strategies and the strategies of other agents. In motion coordination problems, for example, an agent’s utility function typically depends on its position relative to other agents or environmental targets, and knowledge of this function guides local motion adjustments.

In other situations, agents may know nothing about the structure of their utility functions, and how their own utility depends on the actions of other agents (whether local or far away). In this case the only thing they can do is observe rewards based on experience and “optimize” on a trial and error basis. The situation is further complicated because all agents are trying simultaneously to optimize their own strategies. Therefore, even in the absence of noise, an agent trying the same strategy twice may see different results because of the non-stationary nature of the strategies of other agents.

There are several examples of multi-agent systems that illustrate this situation. In distributed routing for ad hoc data networks (e.g., [BK03]), routing nodes seek to route packets to neighboring nodes based on packet destinations without knowledge of the overall network structure. The objective is to minimize the delay of packets to their destinations. This delay must be realized through trial and error, since the functional dependence of delay on routing strategies is not known. A similar problem is automotive traffic routing, in which drives seek to minimize the congestion experienced to get to a desired destination. Drivers can experience the congestion on selected routes as a function of the routes selected by other drivers, but drivers do not know the structure of the congestion function. Finally, in a multi-agent approach to designing manufacturing systems (e.g., [Ger94]), it may not be known in advance how performance measures (such as throughput) depend on manufacturing policy. Rather performance can only be measured once a policy is implemented.

Our interest in this chapter is to develop algorithms that enable coordination in multi-agent systems for precisely this "payoff based" scenario, in which agents only have access to (possibly noisy) measurements of the rewards received through repeated interactions with other agents. We adopt the framework of "learning in games" (see [FL98, Har05, You98, You05] for an extensive overview). Unlike most of the learning rules in this literature, which assume that agents adjust their behavior based on the observed behavior of other agents, we shall assume that agents know only their own past actions and the payoffs that resulted. It is far from obvious that Nash equilibrium can be achieved under such a restriction, but in fact it has recently been shown that such "payoff based" learning rules can be constructed that work in any game [FY06, GL].

In this chapter we show that there are simpler and more intuitive adjustment rules that achieve this objective for a large class of multi-player games known as "weakly acyclic" games. This class captures many problems of interest in cooperative control [MAS07a, MAS07b]. It includes the very special case of "identical interest" games, where each agent receives the same reward. However, weakly acyclic games (and the related concept of potential games) capture other scenarios such as congestion games [Ros73] and similar problems such as distributed routing in networks, weapon target assignment, consensus, and area coverage. See [MAS05, AMS07] and referenced therein for a discussion of a learning in games approach to cooperative control problems, but under less stringent assumptions on informational constraints considered in this chapter.

For many multi-agent problems, operation at a pure Nash equilibrium may reflect optimization of a collective objective.[1] We will derive payoff based dynamics that guarantee asymptotically that agent strategies will constitute a pure Nash equilibrium with arbitrarily high probability. It need not always be the case that at least one Nash equilibrium optimizes a collective objective. Motivated by this consideration, we also discuss the introduction of incentives or tolls in a player's payoff function to assure that there is at least one Nash equilibrium that optimizes a collective objective. Even in this case, however, there may still be suboptimal Nash equilibria.

The remainder of this chapter is organized as follows. Section 5.2 introduces three types of payoff based dynamics in for increasingly general problems. Section 5.2.1 presents "Safe Experimentation Dynamics" which is restricted to identical interest games. Section 5.2.2 presents "Simple Experimentation Dynamics" for the more general class of weakly acyclic games but with noise free payoff measurements. Section 5.2.3 presents "Sample Experimentation Dynamics" for weakly acyclic games with noisy payoff measurements. Section 5.3 discusses how to introduce tolls and

[1] Nonetheless, there are varied viewpoints on the role of Nash equilibrium as a solution concept for multi-agent systems. See [SPG07] and [MS07].

incentives in payoffs so that a Nash equilibrium optimizes a collective objective. Section 5.4 presents an illustrative example of a traffic congestion game. Finally, Section 5.5 contains some concluding remarks. An important analytical tool throughout is the method of resistance trees for perturbed Markov chains [You93], which is reviewed in the appendix of this chapter.

5.2 Payoff Based Learning Algorithms

In this section, we will introduce three simple payoff based learning algorithms. The first, called *Safe Experimentation*, guarantees convergence to a pure optimal Nash equilibrium in any identical interest game. Such an equilibrium is optimal because each player's utility is maximized. The second learning algorithm, called *Simple Experimentation*, guarantees convergence to a pure Nash equilibrium in any weakly acyclic game. The third learning algorithm, called *Sample Experimentation*, guarantees convergence to a pure Nash equilibrium in any weakly acyclic game even when utility measurements are corrupted with noise.

For each learning algorithm, we consider a repeated strategic form game, as described in Section 2.4, with n-player set $\mathcal{P} := \{\mathcal{P}_1, ..., \mathcal{P}_n\}$, a finite action set $\mathcal{A}_i$ for each player $\mathcal{P}_i \in \mathcal{P}$, and a utility function $U_i : \mathcal{A} \to \mathbb{R}$ for each player $\mathcal{P}_i \in \mathcal{P}$, where $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$.

5.2.1 Safe Experimentation Dynamics for Identical Interest Games

5.2.1.1 Constant Exploration Rates

Before introducing the learning dynamics, we introduce the following function. Let

$$U_i^{\max}(t) := \max_{0 \le \tau \le t-1} U_i(a(\tau))$$

be the maximum utility that player $\mathcal{P}_i$ has received up to time $t-1$.

We will now introduce the Safe Experimentation dynamics for identical interest games; see Section 2.3.1 for a review of identical interest games.

1. **Initialization:** At time $t = 0$, each player randomly selects and plays any action, $a_i(0)$. This action will be initially set as the player's *baseline action* at time $t = 1$ and is denoted by $a_i^b(1) = a_i(0)$.

2. **Action Selection:** At each subsequent time step, each player selects his baseline action with probability $(1 - \epsilon)$ or experiments with a new random action with probability ϵ, i.e.:

    - $a_i(t) = a_i^b(t)$ with probability $(1 - \epsilon)$
    - $a_i(t)$ is chosen randomly (uniformly) over a_i with probability ϵ

    The variable ϵ will be referred to as the player's *exploration rate*.

3. **Baseline Strategy Update:** Each player compares the actual utility received, $U_i(a(t))$, with the maximum received utility $U_i^{\max}(t)$ and updates his baseline action as follows:

$$a_i^b(t+1) = \begin{cases} a_i(t), & U_i(a(t)) > U_i^{\max}(t); \\ a_i^b(t), & U_i(a(t)) \leq U_i^{\max}(t). \end{cases}$$

    This step is performed whether or not Step 2 involved exploration.

4. Return to Step 2 and repeat.

The reason that this learning algorithm is called "Safe" Experimentation is that the utility evaluated at the baseline action, $U(a^b(t))$, is non-decreasing with respect to time.

Theorem 5.2.1. *Let G be a finite n-player identical interest game in which all players use the Safe Experimentation dynamics. Given any probability $p < 1$, if the exploration rate $\epsilon > 0$ is sufficiently small, then for all sufficiently large times t, $a(t)$ is an optimal Nash equilibrium of G with at least probability p.*

Proof. Since G is an identical interest game, let the utility of each player be expressed as $U : \mathcal{A} \to \mathbb{R}$ and let $\mathcal{A}^*$ be the set of "optimal" Nash equilibrium of G, i.e.,

$$\mathcal{A}^* = \{a^* \in \mathcal{A} : U(a^*) = \max_{a \in \mathcal{A}} U(a)\}.$$

For any joint action, $a(t)$, the ensuing joint action will constitute an optimal Nash equilibrium with at least probability

$$\left(\frac{\epsilon}{|\mathcal{A}_1|}\right)\left(\frac{\epsilon}{|\mathcal{A}_2|}\right) \cdots \left(\frac{\epsilon}{|\mathcal{A}_n|}\right),$$

where $|\mathcal{A}_i|$ denotes the cardinality of the action set of player $\mathcal{P}_i$. Therefore, an optimal Nash equilibrium will eventually be played with probability 1 for any $\epsilon > 0$.

Suppose an optimal Nash equilibrium is first played at time t^*, i.e., $a(t^*) \in \mathcal{A}^*$ and $a(t^* - 1) \notin \mathcal{A}^*$. Then the baseline joint action must remain constant from that time onwards, i.e., $a^b(t) = a(t^*)$ for all $t > t^*$. An optimal Nash equilibrium will then be played at any time $t > t^*$ with at least probability $(1 - \epsilon)^n$. Since $\epsilon > 0$ can be chosen arbitrarily small, and in particular such that $(1 - \epsilon)^n > p$ this completes the proof. □

5.2.1.2 Diminishing Exploration Rates

In the Safe Experimentation dynamics, the exploration rate ϵ was defined as a constant. Alternatively, one could let the exploration rate vary to induce desirable behavior. One example would be to let the exploration rate decay, such as $\epsilon_t = (1/t)^{1/n}$. This would induce exploration at early stages and reduce exploration at later stages of the game.

The theorem and proof hold under the following conditions for the exploration rate:

$$\lim_{t\to\infty} \epsilon_t = 0,$$

$$\lim_{t\to\infty} \prod_{\tau=1}^{t} \left[1 - \left(\frac{\epsilon_\tau}{|\mathcal{A}_1|}\right)\left(\frac{\epsilon_\tau}{|\mathcal{A}_2|}\right) \cdots \left(\frac{\epsilon_\tau}{|\mathcal{A}_n|}\right) \right] = 0.$$

5.2.2 Simple Experimentation Dynamics for Weakly Acyclic Games

We will now introduce the Simple Experimentation dynamics for weakly acyclic games; see Section 2.3.4 for a review of weakly acyclic games. These dynamics will allow us to relax the assumption of identical interest games.

1. **Initialization:** At time $t = 0$, each player randomly selects and plays any action, $a_i(0)$. This action will be initially set as the player's *baseline action* at time 1, i.e., $a_i^b(1) = a_i(0)$. Likewise, the player's *baseline utility* at time 1 is initialized as $u_i^b(1) = U_i(a(0))$.

2. **Action Selection:** At each subsequent time step, each player selects his baseline action with probability $(1 - \epsilon)$ or experiments with a new random action with probability ϵ.

   - $a_i(t) = a_i^b(t)$ with probability $(1 - \epsilon)$
   - $a_i(t)$ is chosen randomly (uniformly) over a_i with probability ϵ

   The variable ϵ will be referred to as the player's *exploration rate*. Whenever $a_i(t) \neq a_i^b(t)$, we will say that player $\mathcal{P}_i$ *experimented*.

3. **Baseline Action and Baseline Utility Update:** Each player compares the utility received, $U_i(a(t))$, with his baseline utility, $u_i^b(t)$, and updates his baseline action and utility as follows:

   - If player $\mathcal{P}_i$ *experimented* (i.e., $a_i(t) \neq a_i^b(t)$) and if $U_i(a(t)) > u_i^b(t)$ then

$$a_i^b(t+1) = a_i(t),$$

$$u_i^b(t+1) = U_i(a(t)).$$

- If player $\mathcal{P}_i$ *experimented* and if $U_i(a(t)) \leq u_i^b(t)$ then

$$a_i^b(t+1) = a_i^b(t),$$

$$u_i^b(t+1) = u_i^b(t).$$

- If player $\mathcal{P}_i$ *did not experiment* (i.e., $a_i(t) = a_i^b(t)$) then

$$a_i^b(t+1) = a_i^b(t),$$

$$u_i^b(t+1) = U_i(a(t)).$$

4. Return to Step 2 and repeat.

As before, these dynamics require only utility measurements, and hence almost no information regarding the structure of the game.

Theorem 5.2.2. *Let G be a finite n-player weakly acyclic game in which all players use the Simple Experimentation dynamics. Given any probability $p < 1$, if the exploration rate $\epsilon > 0$ is sufficiently small, then for all sufficiently large times t, $a(t)$ is a Nash equilibrium of G with at least probability p.*

The remainder of this subsection is devoted to the proof of Theorem 5.2.2. The proof rely on the theory of resistance trees for perturbed Markov chains (see the appendix of this chapter for a brief review).

Define the *state* of the dynamics to be the pair $[a, u]$, where a is the baseline joint action and u is the baseline utility vector. We will omit the superscript b to avoid cumbersome notation.

Partition the state space into the following three sets. First, let X be the set of states $[a, u]$ such that $u_i \neq U_i(a)$ for at least one player $\mathcal{P}_i$. Let E be the set of states $[a, u]$ such that $u_i = U_i(a)$ for all players $\mathcal{P}_i$ and a is a Nash equilibrium. Let D be the set of states $[a, u]$ such that $u_i = U_i(a)$ for all players $\mathcal{P}_i$ and a is a disequilibrium (not a Nash equilibrium). These are all the states.

Claim 5.2.1. *a. Any state $[a, u] \in X$ transitions to a state in $E \cup D$ in one period with probability $O(1)$.*

b. Any state $[a, u] \in E \cup D$ transitions to a different state $[a', u']$ with probability at most $O(\varepsilon)$.

Proof. For any $[a, u'] \in X$, there exists at least one player $\mathcal{P}_i$ such that $u'_i \neq U_i(a)$. If all players repeat their part of the joint action profile a which occurs with probability $(1 - \epsilon)^n$, then $[a, u']$ transitions to $[a, u]$, where $u_i = U_i(a)$ for all players $\mathcal{P}_i$. Thus the process moves to $[a, u] \in E \cup D$ with prob $O(1)$. This proves statement (a). As for statement (b), any state in $E \cup D$ transitions back to itself whenever no player experiments, which occurs with probability at least $O(1)$. □

Claim 5.2.2. *For any state $[a, u] \in D$, there is a finite sequence of transitions to a state $[a^*, u^*] \in E$, where the transitions have the form*[2]*:*

$$[a, u] \underset{O(\epsilon)}{\rightarrow} [a^1, u^1] \underset{O(\epsilon)}{\rightarrow} \dots \underset{O(\epsilon)}{\rightarrow} [a^*, u^*]$$

where $u_i^k = U_i(a^k)$ for all i and for all $k > 0$, and each transition occurs with probability $O(\epsilon)$.

Proof. Such a sequence is guaranteed by weak acyclicity. Since a is not an equilibrium, there is a better reply path from a to some equilibrium a^*, say $a, a^1, a^2, \dots, a^*$.

At $[a, u]$ the appropriate player $\mathcal{P}_i$ experiments with probability ϵ, chooses the appropriate better reply with probability $1/|\mathcal{A}_i|$, and no one else experiments. Thus the process moves to $[a^1, u^1]$ where $u_i^1 = U_i(a^1)$ for all players $\mathcal{P}_i$ with probability $O(\epsilon)$.

[2] *We will use the notation $z \rightarrow z'$ to denote the transition from state z to state z'. We use $z \underset{O(\epsilon)}{\rightarrow} z'$ to emphasize that this transition occurs with probability of order ϵ.*

Notice that for the deviator $\mathcal{P}_i$, $U_i(a^1) > U_i(a)$, therefore $u_i^1 = U_i(a^1)$. For the non-deviator, say player $\mathcal{P}_j$, $u_j^1 = U_j(a^1)$ since $a_j^1 = a_j$. Thus $[a^1, u^1] \in D \cup E$. In the next period, the appropriate player deviates and so forth.

□

Claim 5.2.3. *For any equilibrium $[a^*, u^*] \in E$, any path from $[a^*, u^*]$ to another state $[a, u] \in E \cup D$, $a \neq a^*$, that does not loop back to $[a^*, u^*]$ must be of one of the following two forms:*

1. $[a^*, u^*] \underset{O(\epsilon)}{\rightarrow} [a^*, u'] \underset{O(\epsilon^k)}{\rightarrow} [a', u''] \rightarrow \ldots \rightarrow [a, u]$*, where* $k \geq 2$*;*
2. $[a^*, u^*] \underset{O(\epsilon^k)}{\rightarrow} [a', u''] \rightarrow \ldots \rightarrow [a, u]$*, where* $k \geq 2$*.*

Proof. The path must begin by either one player experimenting or more that one player experimenting. Case (2) results if more than one player experiments. Case (1) results if exactly one agent, say agent $\mathcal{P}_i$, experiments with an action $a_i' \neq a_i^*$ and all other players continue to play their part of a^*. This happens with probability $\frac{\epsilon}{|\mathcal{A}_i|}(1-\epsilon)^{n-1}$. In this situation, player $\mathcal{P}_i$ cannot be better off, meaning that $U_i(a_i', a_{-i}^*) \leq U_i(a^*)$, since by assumption a^* is an equilibrium. Hence the baseline action next period remains a^* for all players, though their baseline utilities may change. Denote the next state by $[a^*, u']$. If in the subsequent period all players continue to play their part of the action a^* again, which occurs with probability $(1-\epsilon)^n$, then the state reverts back to $[a^*, u^*]$ and we have a loop. Hence the only way the path can continue without a loop is for one or more players to experiment in the next stage, which has probability $O(\epsilon^k)$, $k \geq 1$. This is exactly what case (1) alleges.

□

Proof of Theorem 5.2.2. This is a finite aperiodic Markov process on the state space $\mathcal{A} \times \bar{U}$, where $\bar{U}$ denotes the finite set of baseline utility vectors. Furthermore, from every state there exists a positive probability path to a Nash equilibrium. Hence, every recurrent class has at least one Nash equilibrium. We will now show that within any recurrent class, the trees (see the appendix of this chapter) rooted at the Nash equilibrium will have the lowest resistance. Therefore, according to Theorem 5.6.1, the *a priori* probability that the state will be a Nash equilibrium can be made arbitrarily close to 1.

In order to apply Theorem 5.6.1, we will construct minimum resistance trees with vertices consisting of every possible state (within a recurrence class). Each edge will have resistance $0, 1, 2, ...$ associated with the transition probabilities $O(1), O(\epsilon), O(\epsilon^2), ...$, respectively.

Our analysis will deviate slightly from the presentation in the appendix. In the discussion in the appendix, the vertices of minimum resistance trees are recurrence classes of an associated unperturbed Markov chain. In this case, the unperturbed Markov chain corresponds to Simple Experimentation dynamics with $\epsilon = 0$, and so the recurrence classes are all states in $E \cup D$. Nonetheless, we will construct resistance trees with the vertices being all possible states, i.e., $E \cup D \cup X$. The resulting conclusions remain the same. Since the states in X are transient with probability $O(1)$, the resistance to leave a node corresponding to a state in X is zero. Therefore, the presence of such states does not affect the conclusions determining which states are stochastically stable.

Suppose a minimum resistance tree T is rooted at a vertex v that is not in E. If $v \in X$, it is easy to construct a new tree that has lower resistance. Namely, by Claim 5.2.1a, there is a 0-resistance one-hop path P from v to some state $[a, u] \in E \cup D$. Add the edge of P to T and subtract the edge in T that exits from the vertex $[a, u]$. This results in a $[a, u]$-tree T'. It has lower resistance than T because the added edge has zero resistance while the subtracted edge has resistance greater than or equal to 1 because of Claim 5.2.1b. This argument is illustrated in Figure 5.1, where the red edge of strictly positive resistance is removed and replaced with the blue edge of zero resistance.

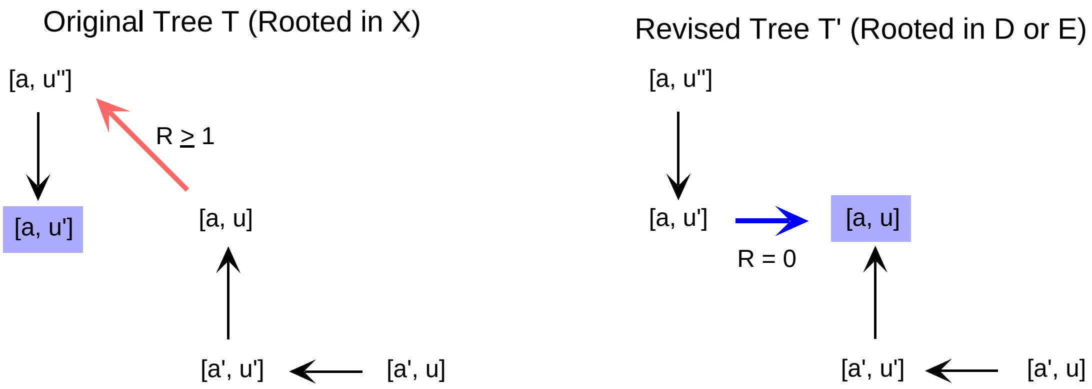


Figure 5.1: Construction of alternative to tree rooted in X.

Suppose next that $v = [a, u] \in D$ but not in E. Construct a path P as in Claim 5.2.2 from $[a, u]$ to some state $[a^*, u^*] \in E$. As above, construct a new tree T' rooted at $[a^*, u^*]$ by adding the edges of P to T and taking out the redundant edges (the edges in T that exit from the vertices in P). The nature of the path P guarantees that the edges taken out have total resistance at least as high as the resistances of the edges put in. This is because the entire path P lies in $E \cup D$, each transition on the path has resistance 1, and, from Claim 5.2.2b, the resistance to leave any state in $E \cup D$ is at least 1.

To construct a new tree that has strictly lower resistance, we will inspect the effect of removing the exiting edge from $[a^*, u^*]$ in T. Note that this edge must fit either case (1) or case (2) of Claim 5.2.3.

In case (2), the resistance of the exiting edge is at least 2, which is larger than any edge in P. Hence the new tree has strictly lower resistance than T, which is a contradiction. This argument is illustrated in Figure 5.2. A new path is created from the original root $[a, u] \in D$ to the equilibrium $[a^*, u^*] \in E$ (blue edges). Redundant (red) edges emanating from the new path are removed. In case (2), the redundant edge emanating from $[a^*, u^*]$ has a resistance of at least 2.

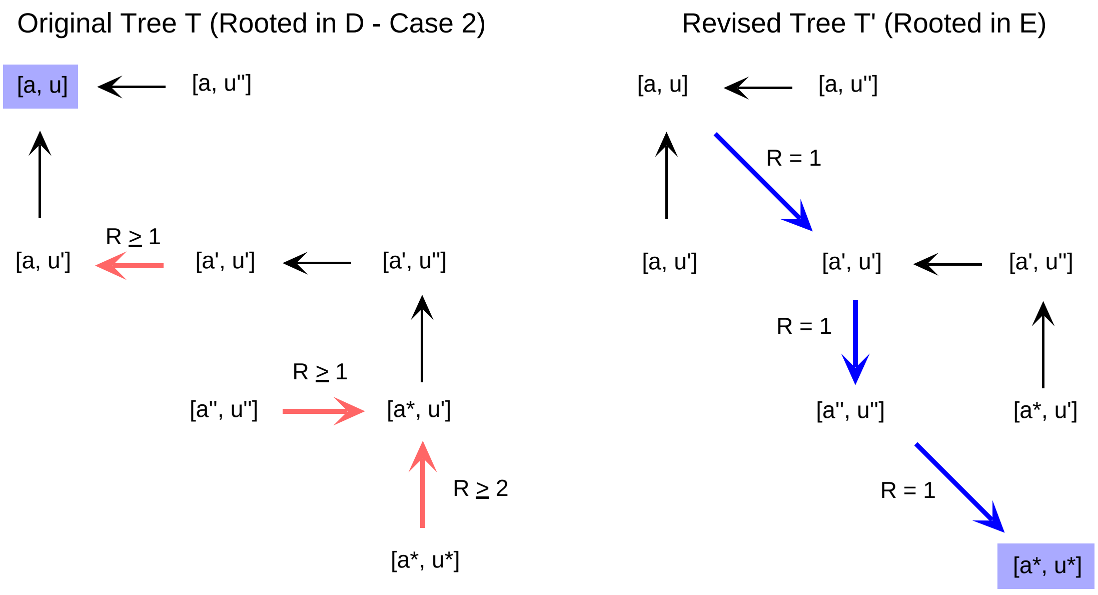


Figure 5.2: Construction of alternative to tree rooted in D for Case (2).

In case (1), the exiting edge has the form $[a^*, u^*] \to [a^*, u']$ which has resistance 1 where $u^* \neq u'$. The next edge in T, say $[a^*, u'] \to [a', u'']$, also has at least resistance 1. Remove the edge $[a^*, u'] \to [a', u'']$ from T, and put in the edge $[a^*, u'] \to [a^*, u^*]$. The latter has resistance 0 since $[a^*, u'] \in X$. This results in a tree T'' that is rooted at $[a^*, u^*]$ and has strictly lower resistance than does T, which is a contradiction. This argument is illustrated in Figure 5.3. As in Figure 5.2, a new (blue) path is constructed and redundant (red) edges are removed. The difference is that the edge $[a^*, u'] \to [a', u'']$ is removed and replaced with $[a^*, u'] \to [a^*, u^*]$.

To recap, a minimum resistant tree cannot be rooted at any state in X or D, and therefore can only be rooted in E. Therefore, when ϵ is sufficiently small, the long-run probability on E can be made arbitrarily close to 1, and in particular larger than any specified probability p. □

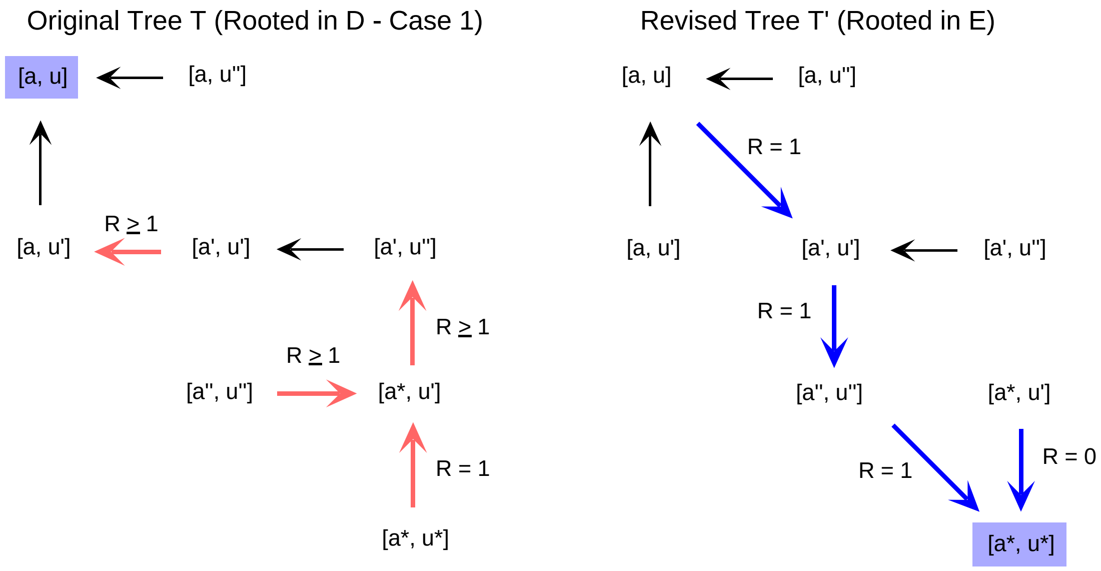


Figure 5.3: Construction of alternative to tree rooted in D for Case (1).

5.2.3 Sample Experimentation Dynamics for Weakly Acyclic Games with Noisy Utility Measurements

5.2.3.1 Noise-free Utility Measurements

In this section we will focus on developing payoff based dynamics for which the limiting behavior exhibits that of a pure Nash equilibrium with arbitrarily high probability in any finite weakly acyclic game *even in the presence of utility noise*. We will show that a variant of the so-called Regret Testing algorithm [FY06] accomplishes this objective for weakly acyclic games with noisy utility measurements.

We now introduce Sample Experimentation dynamics.

1. **Initialization:** At time $t = 0$, each player randomly selects and plays any action, $a_i(0) \in \mathcal{A}_i$. This action will be initially set as the player's *baseline action*, $a_i^b(1) = a_i(0)$.

2. **Exploration Phase:** After the baseline action is set, each player engages in an

exploration phase over the next m periods. The length of the exploration phase need not be the same or synchronized for each player, but we will assume that they are for the proof. For convenience, we will double index the time of the actions played as

$$\breve{a}(t_1, t_2) = a(m\, t_1 + t_2)$$

where t_1 indexes the number of the exploration phase and t_2 indexes the actions played in that exploration phase. We will refer to t_1 as the *exploration phase time* and t_2 as the *exploration action time*. By construction, the exploration phase time and exploration action time satisfy $t_1 \geq 1$ and $m \geq t_2 \geq 1$. The baseline action will only be updated at the end of the exploration phase and will therefore only be indexed by the exploration phase time.

During the exploration phase, each player selects his baseline action with probability $(1 - \epsilon)$ or experiments with a new random action with probability ϵ. That is, for any exploration phase time $t_1 \geq 1$ and for any exploration action time satisfying $m \geq t_2 \geq 1$,

- $\breve{a}_i(t_1, t_2) = a_i^b(t_1)$ with probability $(1 - \epsilon)$,
- $\breve{a}_i(t_1, t_2)$ is chosen randomly (uniformly) over $(\mathcal{A}_i \backslash a_i^b(t_1))$ with probability ϵ.

Again, the variable ϵ will be referred to as the player's *exploration rate*.

3. **Action Assessment:** After the exploration phase, each player evaluates the average utility received when playing each of his actions during the exploration phase. Let $n_i^{a_i}(t_1)$ be the number of times that player $\mathcal{P}_i$ played action a_i during the exploration phase at time t_1. The average utility for action a_i during the

exploration phase at time t_1 is

$$\hat{V}_i^{a_i}(t_1) = \begin{cases} \frac{1}{n_i^{a_i}(t_1)} \sum_{t_2=1}^{m} I\{a_i = \check{a}_i(t_1, t_2)\} U_i(\check{a}(t_1, t_2)), & n_i^{a_i}(t_1) > 0; \\ U_{\min}, & n_i^{a_i}(t_1) = 0, \end{cases}$$

where $I\{\cdot\}$ is the usual indicator function and U_{min} satisfies

$$U_{\min} < \min_i \min_{a \in \mathcal{A}} U_i(a).$$

In words, $U_{\min}$ is less than the smallest payoff any agent can receive.

4. **Evaluation of Better Response Set:** Each player compares the average utility received when playing his baseline action, $\hat{V}_i^{a_i^b(t)}(t_1)$, with the average utility received for each of his other actions, $\hat{V}_i^{a_i}(t_1)$, and finds all played actions which performed δ better than the baseline action. The term δ will be referred to as the players' *tolerance level*. Define $\mathcal{A}_i^*(t_1)$ to be the set of actions that outperformed the baseline action as follows:

$$\mathcal{A}_i^*(t_1) := \left\{ a_i \in \mathcal{A}_i : \hat{V}_i^{a_i}(t_1) \geq \hat{V}_i^{a_i^b(t_1)}(t_1) + \delta \right\}. \tag{5.1}$$

5. **Baseline Strategy Update:** Each player updates his baseline action as follows:

   - If $\mathcal{A}_i^*(t_1) = \emptyset$, then $a_i^b(t_1 + 1) = a_i^b(t_1)$.
   - If $\mathcal{A}_i^*(t_1) \neq \emptyset$, then
     - With probability ω, set $a_i^b(t_1 + 1) = a_i^b(t_1)$. (We will refer to ω as the player's inertia.)
     - With probability $1 - \omega$, randomly select $a_i^b(t_1 + 1) \in \mathcal{A}_i^*(t_1)$ with uniform probability.

6. Return to Step 2 and repeat.

For simplicity, we will first state and prove the desired convergence properties using noiseless utility measurements. The setup for the noisy utility measurements will be stated afterwards.

Before stating the following theorem, we define the constant $\alpha > 0$ as follows. If $U_i(a^1) \neq U_i(a^2)$ for any joint actions $a^1, a^2 \in \mathcal{A}$ and any player $\mathcal{P}_i \in \mathcal{P}$, then $|U_i(a^1) - U_i(a^2)| > \alpha$. In words, if any two joint actions result in different utilities at all, then the difference would be at least α.

Theorem 5.2.3. *Let G be a finite n-player weakly acyclic game in which all players use the Sample Experimentation dynamics. For any*

- *probability $p < 1$,*
- *tolerance level $\delta \in (0, \alpha)$,*
- *inertia $\omega \in (0, 1)$, and*
- *exploration rate ϵ satisfying $\min\{(\alpha - \delta)/4, \delta/4, 1 - p\} > (1 - (1 - \epsilon)^n) > 0$,*

if the exploration phase length m is sufficiently large, then for all sufficiently large times $t > 0$, $a(t)$ is a Nash equilibrium of G with at least probability p.

The remainder of this subsection is devoted to the proof of Theorem 5.2.3.

We will assume for simplicity that utilities are between -1/2 and 1/2, i.e., $|U_i(a)| \leq 1/2$ for any player $\mathcal{P}_i \in \mathcal{P}$ and any joint action $a \in \mathcal{A}$.

We begin with a series of useful claims. The first claim states that for any player $\mathcal{P}_i$ the average utility for an action $a_i \in \mathcal{A}_i$ during the exploration phase can be made arbitrarily close (with high probability) to the actual utility the player would have received provided that all other players never experimented. This can be accomplished if the experimentation rate is sufficiently small and the exploration phase length is sufficiently large.

Claim 5.2.4. *Let a^b be the joint baseline action at the start of an exploration phase of length m. For*

- *any probability $p < 1$,*
- *any $\delta^* > 0$, and*
- *any exploration rate $\epsilon > 0$ satisfying $\delta^*/2 \geq (1 - (1-\epsilon)^{n-1}) > 0$,*

if the exploration phase length m is sufficiently large then

$$\mathbf{Pr}\left[\left|\hat{V}_i^{a_i} - U_i(a_i, a_{-i}^b)\right| > \delta^*\right] < 1 - p.$$

Proof. Let $n_i(a_i)$ represent the number of times player $\mathcal{P}_i$ played action a_i during the exploration phase. In the following discussion, *all probabilities and expectations are conditioned on* $n_i(a_i) > 0$. We omit making this explicit for the sake of notational simplicity. The event $n_i(a_i) = 0$ has diminishing probability as the exploration phase length m increases, and so this case will not affect the desired conclusions for increasing phase lengths.

For an arbitrary $\delta^* > 0$,

$$\begin{aligned}
&\mathbf{Pr}\left[\left|\hat{V}_i^{a_i} - U_i(a_i, a_{-i}^b)\right| > \delta^*\right] \\
&\quad \leq \mathbf{Pr}\left[\left|\hat{V}_i^{a_i} - E\{\hat{V}_i^{a_i}\}\right| + \left|E\{\hat{V}_i^{a_i}\} - U_i(a_i, a_{-i}^b)\right| > \delta^*\right] \\
&\quad \leq \underbrace{\mathbf{Pr}\left[\left|\hat{V}_i^{a_i} - E\{\hat{V}_i^{a_i}\}\right| > \delta^*/2\right]}_{(*)} + \underbrace{\mathbf{Pr}\left[\left|E\{\hat{V}_i^{a_i}\} - U_i(a_i, a_{-i}^b)\right| > \delta^*/2\right]}_{(**)}.
\end{aligned}$$

First, let us focus on $(**)$. We have

$$E\{\hat{V}_i^{a_i}\} - U_i(a_i, a_{-i}^b) = [1 - (1-\epsilon)^{n-1}]\left[E\{U_i(a_i, a_{-i}(t))|a_{-i}(t) \neq a_{-i}^b\} - U_i(a_i, a^b)\right],$$

which approaches 0 as $\epsilon \downarrow 0$. Therefore, for any exploration rate ϵ satisfying $\delta^*/2 > (1-(1-\epsilon)^{n-1}) > 0$, we know that

$$\mathbf{Pr}\left[\left|E\{\hat{V}_i^{a_i}\} - U_i(a_i, a_{-i}^b)\right| > \delta^*/2\right] = 0.$$

Now we will focus on $(*)$. By the weak law of large numbers, $(*)$ approaches 0 as $n_i(a_i) \uparrow \infty$. This implies that for any probability $\bar{p} < 1$ and any exploration rate $\epsilon > 0$, there exists a sample size $n_i^*(a_i)$ such that if $n_i(a_i) > n_i^*(a_i)$ then

$$\mathbf{Pr}\left[\left|\hat{V}_i^{a_i} - E\{\hat{V}_i^{a_i}\}\right| > \rho/2\right] < 1 - \bar{p}.$$

Lastly, for any probability $\bar{p} < 1$ and any fixed exploration rate, there exists a minimum exploration length $\underline{m} > 0$ such that for any exploration length $m > \underline{m}$,

$$\mathbf{Pr}\left[n_i(a_i) \geq n_i^*(a_i)\right] \geq \bar{p}.$$

In summary, for any fixed exploration rate ϵ satisfying $\delta^*/2 \geq (1-(1-\epsilon)^{n-1}) > 0$, $(*) + (**)$ can be made arbitrarily close to 0, provided that the exploration length m is sufficiently large. □

Claim 5.2.5. *Let a^b be the joint baseline action at the start of an exploration phase of length m. For any*

- *probability $p < 1$,*
- *tolerance level $\delta \in (0, \alpha)$, and*
- *exploration rate $\epsilon > 0$ satisfying $\min\{(\alpha-\delta)/4, \delta/4\} \geq (1-(1-\epsilon)^{n-1}) > 0$,*

if the exploration length m is sufficiently large, then each player's better response set a_i^* will contain ***only and all*** *actions that are a better response to the joint baseline action, i.e.,*

$$a_i^* \in \mathcal{A}_i^* \Leftrightarrow U_i(a_i^*, a_{-i}^b) > U_i(a^b)$$

with at least probability p.

Proof. Suppose a^b is not a Nash equilibrium. For some player $\mathcal{P}_i \in \mathcal{P}$, let a_i^* be a strict better reply to the baseline joint action, i.e. $U_i(a_i^*, a_{-i}^b) > U_i(a^b)$ and let a_i^w be a non-better reply to the baseline joint action, i.e. $U_i(a_i^w, a_{-i}^b) \leq U_i(a^b)$.

Using Claim 5.2.4, for any probability $\bar{p} < 1$ and any exploration rate $\epsilon > 0$ satisfying $\min\{(\alpha - \delta)/4, \delta/4\} \geq (1 - (1 - \epsilon)^{n-1}) > 0$ there exists a minimum exploration length $\underline{m} > 0$ such that for any exploration length $m > \underline{m}$ the following expressions are true:

$$\mathbf{Pr}\left[|\hat{V}_i^{a_i^b} - U_i(a_i^b, a_{-i}^b)| < \delta^*\right] \geq \bar{p}, \tag{5.2}$$

$$\mathbf{Pr}\left[|\hat{V}_i^{a_i^*} - U_i(a_i^*, a_{-i}^b)| < \delta^*\right] \geq \bar{p}, \tag{5.3}$$

$$\mathbf{Pr}\left[|\hat{V}_i^{a_i^w} - U_i(a_i^w, a_{-i}^b)| < \delta^*\right] \geq \bar{p}, \tag{5.4}$$

where $\delta^* = \min\{(\alpha - \delta)/2, \delta/2\}$. Rewriting equation 5.2 we obtain

$$\mathbf{Pr}\left[|\hat{V}_i^{a_i^b} - U_i(a_i^b, a_{-i}^b)| < \delta^*\right] \leq \mathbf{Pr}\left[\hat{V}_i^{a_i^b} - U_i(a_i^b, a_{-i}^b) < (\alpha - \delta)/2\right],$$

and rewriting equation 5.3 we obtain

$$\begin{aligned}\mathbf{Pr}\left[|\hat{V}_i^{a_i^*} - U_i(a_i^*, a_{-i}^b)| < \delta^*\right] &\leq \mathbf{Pr}\left[\hat{V}_i^{a_i^*} - U_i(a_i^*, a_{-i}^b) > -(\alpha - \delta)/2\right], \\ &\leq \mathbf{Pr}\left[\hat{V}_i^{a_i^*} - (U_i(a_i^b, a_{-i}^b) + \alpha) > -(\alpha - \delta)/2\right], \\ &= \mathbf{Pr}\left[\hat{V}_i^{a_i^*} - U_i(a_i^b, a_{-i}^b) > (\alpha + \delta)/2\right],\end{aligned}$$

meaning that

$$\mathbf{Pr}\left[a_i^* \in \mathcal{A}_i^*\right] \geq \bar{p}^2.$$

Similarly, rewriting equation 5.2 we obtain

$$\mathbf{Pr}\left[|\hat{V}_i^{a_i^b} - U_i(a_i^b, a_{-i}^b)| < \delta^*\right] \leq \mathbf{Pr}\left[\hat{V}_i^{a_i^b} - U_i(a_i^b, a_{-i}^b) > -\delta/2\right],$$

and rewriting equation 5.4 we obtain

$$\begin{aligned}\mathbf{Pr}\left[|\hat{V}_i^{a_i^w} - U_i(a_i^w, a_{-i}^b)| < \delta^*\right] &\leq \mathbf{Pr}\left[\hat{V}_i^{a_i^w} - U_i(a_i^w, a_{-i}^b) < \delta/2\right], \\ &\leq \mathbf{Pr}\left[\hat{V}_i^{a_i^w} - U_i(a_i^b, a_{-i}^b) < \delta/2\right],\end{aligned}$$

meaning that

$$\mathbf{Pr}\left[a_i^w \notin \mathcal{A}_i^*\right] \quad \geq \quad \bar{p}^2.$$

Since $\bar{p}$ can be chosen arbitrarily close to 1, the proof is complete. □

Theorem 5.2.3. The evolution of the baseline actions from phase to phase is a finite aperiodic Markov process on the state space of joint actions, $\mathcal{A}$. Furthermore, since G is weakly acyclic, from every state there exists a better reply path to a Nash equilibrium. Hence, every recurrent class has at least one Nash equilibrium. We will show that these dynamics can be viewed as a perturbation of a certain a Markov chain whose recurrent classes are restricted to Nash equilibria. We will then appeal to Theorem 5.6.1 to derive the desired result.

We begin by defining an "unperturbed" process on baseline actions. For any $a^b \in \mathcal{A}$, define the *true* better reply set as

$$\bar{\mathcal{A}}_i^*(a^b) := \left\{a_i : U_i(a_i, a_{-i}^b) > U_i(a^b)\right\}.$$

Now define the transition process from $a^b(t_1)$ to $a^b(t_1+1)$ as follows:

- If $\bar{\mathcal{A}}_i^*(a^b(t_1)) = \emptyset$, then $a_i^b(t_1+1) = a_i^b(t_1)$.
- If $\bar{\mathcal{A}}_i^*(a^b(t_1)) \neq \emptyset$, then
  - With probability ω, set $a_i^b(t_1+1) = a_i^b(t_1)$.
  - With probability $1-\omega$, randomly select $a_i^b(t_1+1) \in \bar{\mathcal{A}}_i^*(t_1)$ with uniform probability.

This is a special case of a so-called "better reply process with finite memory and inertia". From [You05, Theorem 6.2], the joint actions of this process converge to a Nash equilibrium with probability 1 in any weakly acyclic game. Therefore, the recurrence classes of this unperturbed are precisely the set of pure Nash equilibria.

The above unperturbed process closely resembles the Baseline Strategy Update process described in Step 5 of Sample Experimentation Dynamics. The difference is that the above process uses the true better reply set, whereas Step 5 uses a better reply set constructed from experimentation over a phase. However, by Claim 5.2.5, for any probability $\bar{p} < 1$, acceptable tolerance level δ, and acceptable exploration rate ϵ, there exists a minimum exploration phase length $\underline{m}$ such that for any exploration phase length $m > \underline{m}$, each player's better response set will contain only and all actions that are a strict better response with at least probability $\bar{p}$.

With parameters selected according to Claim 5.2.5, the transitions of the baseline joint actions in Sample Experimentation Dynamics follow that of the above unperturbed better reply process with probability $\bar{p}$ arbitrarily close to 1. Since the recurrence classes of the unperturbed process are only Nash equilibria, we can conclude from Theorem 5.6.1 that as $\bar{p}$ approaches 1, the probability that the baseline action for sufficiently large t_1 will be a (pure) Nash equilibrium can be made arbitrarily close to 1. By selecting the exploration probability ϵ sufficiently small, we can also conclude that the joint action during exploration phases, i.e., $a(mt_1 + t_2)$, will also be a Nash equilibrium with probability arbitrarily close to 1.

□

5.2.3.2 Noisy Utility Measurements

Suppose that each player receives a noisy measurement of his true utility, i.e.,

$$\tilde{U}_i(a_i, a_{-i}) = U_i(a_i, a_{-i}) + \nu_i,$$

where ν_i is an i.i.d. random variable with zero mean. In the regret testing algorithm with noisy utility measurements, the average utility for action a_i during the exploration phase at time t_1 is now

$$\hat{V}_i^{a_i}(t_1) = \begin{cases} \frac{1}{n_i^{a_i}(t_1)} \sum_{t_2=1}^{m} I\{a_i = \check{a}_i(t_1,t_2)\}\tilde{U}_i(\check{a}(t_1,t_2)), & n_i^{a_i}(t_1) > 0; \\ U_{\min}, & n_i^{a_i}(t_1) = 0. \end{cases}$$

A straightforward modification of the proof of Theorem 5.2.3 leads to the following theorem.

Theorem 5.2.4. *Let G be a finite n-player weakly acyclic game where players' utilities are corrupted with a zero mean noise process. If all players use the regret testing dynamics, then for any*

- *probability $p < 1$,*
- *tolerance level $\delta \in (0, \alpha)$,*
- *inertia $\omega \in (0, 1)$, and*
- *exploration rate ϵ satisfying $\min\{(\alpha - \delta)/4, \delta/4, 1 - p\} > (1 - (1 - \epsilon)^n) > 0$,*

if the exploration phase length m is sufficiently large, then for all sufficiently large times $t > 0$, $a(t)$ is a Nash equilibrium of G with at least probability p.

5.2.3.3 Comment on Length and Synchronization of Players' Exploration Phases

In the proof of Theorem 5.2.3, we assumed that all players' exploration phases were synchronized and of the same length. This assumption was used to ensure that the baseline action of the other players remained constant when a player assessed the performance of a particular action. Because of the players' inertia this assumption is unnecessary. The general idea is as follows: a player will repeat his baseline action regardless of his better response set with positive probability because of his inertia. Therefore, if all players repeat their baseline action a sufficient number of times, which happens with positive probability, then the joint baseline action would remain constant long enough for any player to evaluate an accurate better response set for that particular joint baseline action.

5.3 Influencing Nash Equilibria in Resource Allocation Problems

In this section we will derive an approach for influencing the Nash equilibria of a resource allocation problem using the idea of marginal cost pricing. We will illustrate the setup and our approach on a congestion game which is an example of a resource allocation problem.

5.3.1 Congestion Game with Tolls Setup

We consider a congestion game, as defined in Section 2.3.3, with a player set $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_n\}$, a set of resources $\mathcal{R}$, and a congestion $c_r : \{0, 1, 2, ...\} \rightarrow \mathbb{R}$ for each resource $r \in \mathcal{R}$.

One approach for equilibrium manipulation is to influence drivers' utilities with tolls [San02], as introduced in Section 3.4.2. In a congestion game with tolls, a driver's utility takes on the form

$$U_i(a) = -\sum_{r \in \mathcal{A}_i} c_r(\sigma_r(a)) + t_r(\sigma_r(a)),$$

where $t_r(k)$ is the toll imposed on route r if there are k users.

In Section 3.4.2, we analyzed the situation in which a global planner was interested in minimizing the total congestion experienced by all drivers on the network, which can be evaluated

$$T_c(a) := \sum_{r \in R} \sigma_r(a) c_r(\sigma_r(a)).$$

Now suppose that the global planner is interested in minimizing a *more general* measure[3],

$$\phi(a) := \sum_{r\in R} f_r(\sigma_r(a))c_r(\sigma_r(a)). \tag{5.5}$$

An example of an objective function that fits within this framework and may be practical for general resource allocation problems is

$$\phi(a) = \sum_{r\in R} c_r(\sigma_r(a)).$$

We will now show that there exists a set of tolls, $t_r(\cdot)$, such that the potential function associated with the congestion game with tolls will be aligned with the global planner's objective function of the form given in equation (5.5).

Proposition 5.3.1. *Consider a congestion game of any network topology. If the imposed tolls are set as*

$$t_r(k) = (f_r(k) - 1)c_r(k) - f_r(k-1)c_r(k-1), \;\; \forall k \geq 1,$$

then the global planners objective, $\phi_c(a) = -\phi(a)$, is a potential function for the congestion game with tolls.

Proof. Let $a^1 = \{a_i^1, a_{-i}\}$ and $a^2 = \{a_i^2, a_{-i}\}$. We will use the shorthand notation $\sigma_r^{a^1}$ to represent $\sigma_r(a^1)$. The change in utility incurred by driver d_i in changing from route a_i^2 to route a_i^1 is

$$\begin{aligned} U_i(a^1) - U_i(a^2) &= -\sum_{r\in\mathcal{A}_i^1}\big(c_r(\sigma_r^{a^1}) + t_r(\sigma_r^{a^1})\big) + \sum_{r\in\mathcal{A}_i^2}\big(c_r(\sigma_r^{a^2}) + t_r(\sigma_r^{a^2})\big), \\ &= -\sum_{r\in\mathcal{A}_i^1\setminus a_i^2}\big(c_r(\sigma_r^{a^1}) + t_r(\sigma_r^{a^1})\big) + \sum_{r\in\mathcal{A}_i^2\setminus a_i^1}\big(c_r(\sigma_r^{a^2}) + t_r(\sigma_r^{a^2})\big). \end{aligned}$$

[3] In fact, if $c_r(\sigma_r(a)) \neq 0$ for all a, then (5.5) is equivalent to $\sum_{r\in R}\tilde{f}_r(\sigma_r(a))$ where $f_r(\sigma_r(a)) = \frac{\tilde{f}_r(\sigma_r(a))}{c_r(\sigma_r(a))}$.

The change in the total negative congestion from the joint action a^2 to a^1 is

$$\phi_c(a^1) - \phi_c(a^2) \;=\; -\sum_{r \in (a_i^1 \cup a_i^2)} \big(f_r(\sigma_r^{a^1})c_r(\sigma_r^{a^1}) - f_r(\sigma_r^{a^2})c_r(\sigma_r^{a^2})\big).$$

Since

$$\sum_{r \in (a_i^1 \cap a_i^2)} \big(f_r(\sigma_r^{a^1})c_r(\sigma_r^{a^1}) - f_r(\sigma_r^{a^2})c_r(\sigma_r^{a^2})\big) = 0,$$

the change in the total negative congestion is

$$\begin{aligned}\phi_c(a^1) - \phi_c(a^2) = &- \sum_{r \in \mathcal{A}_i^1 \setminus a_i^2} \big(f_r(\sigma_r^{a^1})c_r(\sigma_r^{a^1}) - f_r(\sigma_r^{a^2})c_r(\sigma_r^{a^2})\big)\\ &- \sum_{r \in \mathcal{A}_i^2 \setminus a_i^1} \big(f_r(\sigma_r^{a^1})c_r(\sigma_r^{a^1}) - f_r(\sigma_r^{a^2})c_r(\sigma_r^{a^2})\big).\end{aligned}$$

Expanding the first term, we obtain

$$\begin{aligned}&\sum_{r \in \mathcal{A}_i^1 \setminus a_i^2} \big(f_r(\sigma_r^{a^1})c_r(\sigma_r^{a^1}) - f_r(\sigma_r^{a^2})c_r(\sigma_r^{a^2})\big)\\ &\qquad = \sum_{r \in \mathcal{A}_i^1 \setminus a_i^2} \big(f_r(\sigma_r^{a^1})c_r(\sigma_r^{a^1}) - (f_r(\sigma_r^{a^1} - 1))c_r(\sigma_r^{a^1} - 1)\big),\\ &\qquad = \sum_{r \in \mathcal{A}_i^1 \setminus a_i^2} \big(f_r(\sigma_r^{a^1})c_r(\sigma_r^{a^1}) - ((f_r(\sigma_r^{a^1}) - 1)c_r(\sigma_r^{a^1}) - t_r(\sigma_r^{a^1}))\big),\\ &\qquad = \sum_{r \in \mathcal{A}_i^1 \setminus a_i^2} \big(c_r(\sigma_r^{a^1}) + t_r(\sigma_r^{a^1})\big).\end{aligned}$$

Therefore,

$$\begin{aligned}\phi_c(a^1) - \phi_c(a^2) \;&=\; -\sum_{r \in \mathcal{A}_i^1 \setminus a_i^2} \big(c_r(\sigma_r^{a^1}) + t_r(\sigma_r^{a^1})\big) + \sum_{r \in \mathcal{A}_i^2 \setminus a_i^1} \big(c_r(\sigma_r^{a^2}) + t_r(\sigma_r^{a^2})\big),\\ &=\; U_i(a^1) - U_i(a^2).\end{aligned}$$

□

By implementing the tolling scheme set forth in Proposition 5.3.1, we guarantee that all action profiles that minimize the global planner's objective are equilibrium of the congestion game with tolls.

In the special case that $f_r(\sigma_r(a)) = \sigma_r(a)$, then Proposition 5.3.1 produces the same tolls as in Proposition 3.4.1

5.4 Illustrative Example – Braess' Paradox

We will consider a discrete representation of the congestion game setup considered in Braess' Paradox [Bra68]. In our setting, there are 1000 vehicles that need to traverse through the network. The network topology and associated congestion functions are illustrated in Figure 5.4. Each vehicle can select one of the four possible paths to traverse across the network.

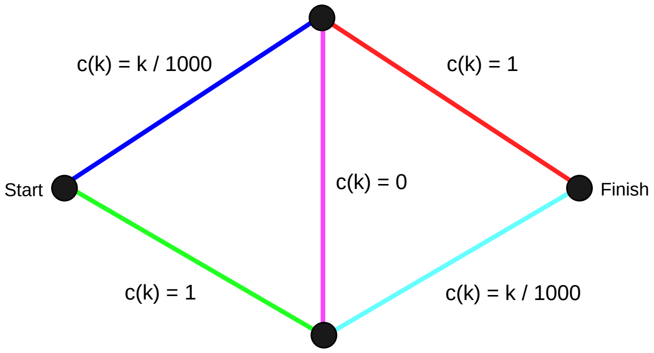


Figure 5.4: Congestion Game Setup – Braess' Paradox

The reason for using this setup as an illustration of the learning algorithms and equilibrium manipulation approach developed in this chapter is that the Nash equilibrium of this particular congestion game is easily identifiable. The unique Nash equilibrium is when all vehicles take the route as highlighted in Figure 5.5. At this Nash equilibrium each vehicle has a utility of 2 and the total congestion is 2000.

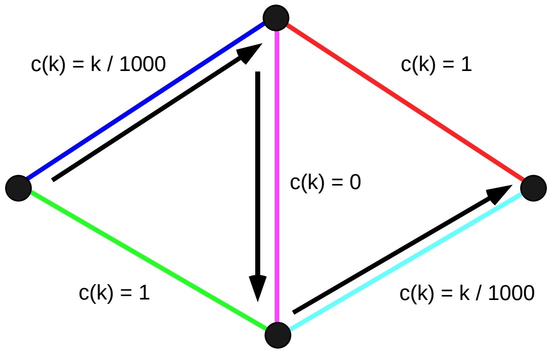


Figure 5.5: Illustration of Nash Equilibrium in Braess' Paradox.

Since a potential game is weakly acyclic, the payoff based learning dynamics in this chapter are applicable learning algorithms for this congestion game. In a congestion game, a payoff based learning algorithms means that drivers have access *only* to the actual congestion experienced. Drivers are unaware of the congestion level on any alternative routes. Figure 5.6 shows the evolution of drivers on routes when using the Simple Experimentation dynamics. This simulation used an experimentation rate of $\epsilon = 0.25\%$. The colors on the plots are consistent with the colors of each route as indicated in Figure 5.4. One can observe that the vehicles' collective behavior does indeed approach that of the Nash equilibrium.

In this congestion game, it is also easy to verify that this vehicle distribution does not minimize the total congestion experience by all drivers over the network. The distribution that minimizes the total congestion over the network is when half the vehicles occupy the top two roads and the other half occupy the bottom two roads. The middle road (pink) is irrelevant.

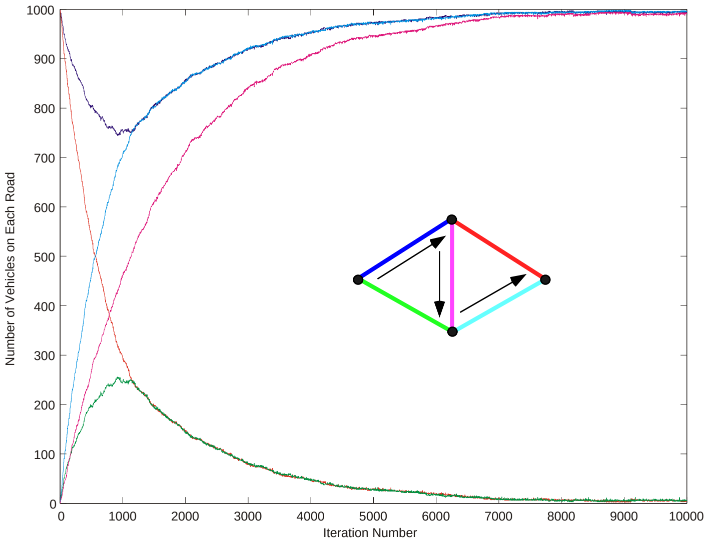


Figure 5.6: Braess' Paradox: Evolution of Number of Vehicles on Each Road Using Simple Experimentation Dynamics

One can employ the tolling scheme developed in the previous section to locally influence vehicle behavior to achieve this objective. In this setting, the new cost functions, i.e. congestion plus tolls, are illustrated in Figure 5.7.

Figure 5.8 shows the evolution of drivers on routes when using the Simple Experimentation dynamics. This simulation used an experimentation rate of $\epsilon = 0.25\%$. When using this tolling scheme, the vehicles' collective behavior approaches the refined Nash equilibrium which now minimizes the total congestion experienced on the network. The total congestion experienced on the network is now approximately 1500.

There are other tolling schemes that would have resulted in the desired allocation. One approach is to assign an infinite cost to the middle road, which is equivalent to removing it from the network. Under this scenario, the unique Nash equilibrium is for half the vehicles to occupy the top route and half the bottom, which would minimize

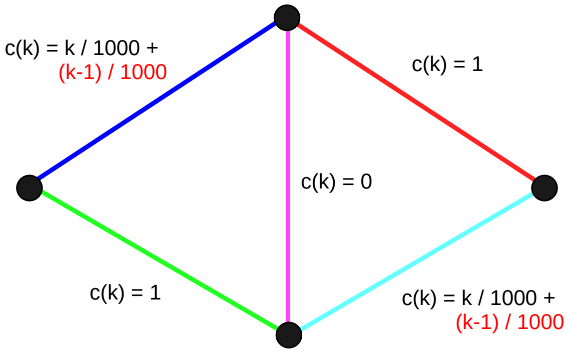


Figure 5.7: Braess' Paradox: Congestion Game Setup with Tolls to Minimize Total Congestion

the total congestion on the network. Therefore, the existence of this extra road, even though it has zero cost, resulted in the unique Nash equilibrium having a higher total congestion. This is Braess' Paradox [Bra68].

The advantage of the tolling scheme set forth in this chapter is that it gives a systematic method for influencing the Nash equilibria of any congestion game. We would like to highlight that this tolling scheme only guarantees that the action profiles that maximize the desired objective function are Nash equilibria of the new congestion game with tolls. However, it does not guarantee the lack of suboptimal Nash equilibria.

In many applications, players may not have access to their true utility, but do have access to a noisy measurement of their utility. For example, in the traffic setting, this noisy measurement could be the result of accidents or weather conditions. We will revisit the original congestion game (without tolls) as illustrated in Figure 5.4. We will now assume that a driver's utility measurement takes on the form

$$\tilde{U}_i(a) = -\sum_{r \in \mathcal{A}_i} c_r(\sigma_r(a)) + \nu_i,$$

where ν_i is a random variable with zero mean and variance of 0.1. We will assume that

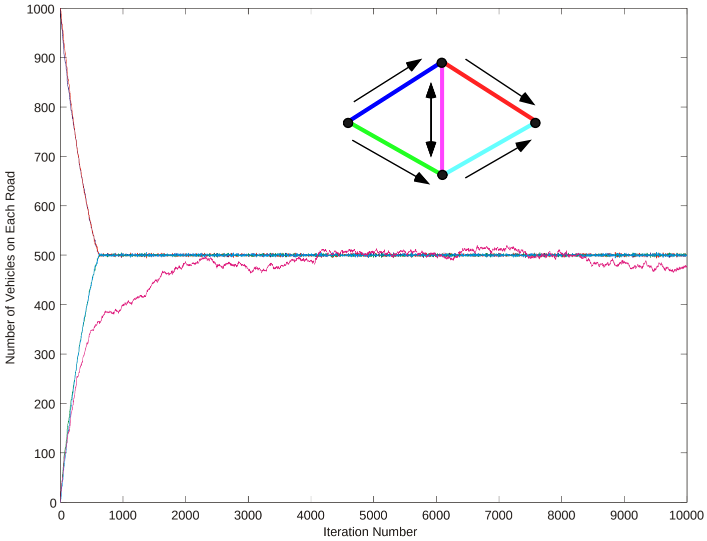


Figure 5.8: Braess' Paradox: Evolution of Number of Vehicles on Each Road Using Simple Experimentation Dynamics with Optimal Tolls

the noise is driver specific rather than road specific.

Figure 5.9 shows a comparison of the evolution of drivers on routes when using the Simple and Sample Experimentation dynamics. The Simple Experimentation dynamics simulation used an experimentation rate $\epsilon = 0.25\%$. The Sample Experimentation dynamics simulation used an exploration rate $\epsilon = 0.25\%$, a tolerance level $\delta = 0.002$, an exploration phase length $m = 500000$, and inertia $\omega = 0.85$. As expected, the noisy utility measurements influenced vehicle behavior more in the Simple Experimentation dynamics than the Sample Experimentation dynamics.

5.5 Concluding Remarks and Future Work

We have introduced Safe Experimentation dynamics for identical interest games, Simple Experimentation dynamics for weakly acyclic games with noise-free utility mea-

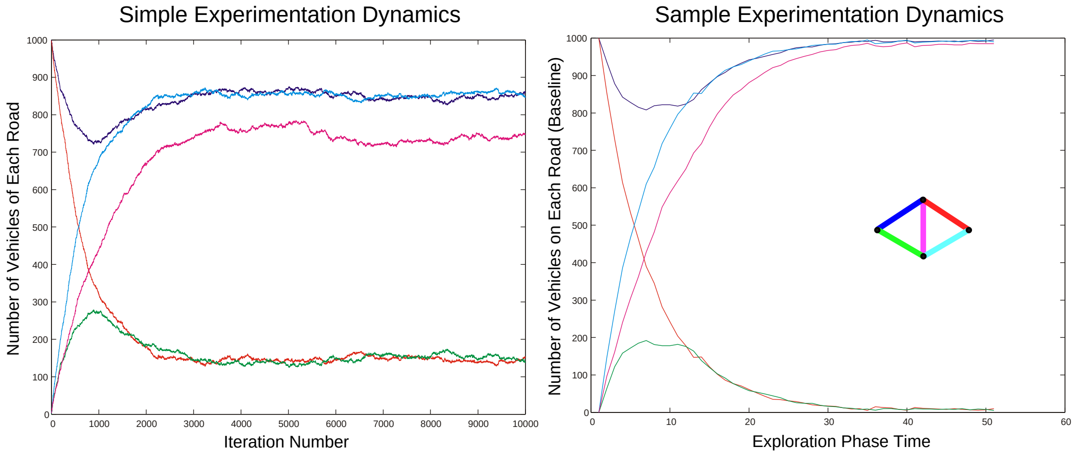


Figure 5.9: Braess' Paradox: Comparison of Evolution of Number of Vehicles on Each Road Using Simple Experimentation Dynamics and Sample Experimentation Dynamics (baseline) with Noisy Utility Measurements

surements, and Sample Experimentation dynamics for weakly acyclic games with noisy utility measurements. For all three settings, we have shown that for sufficiently large times, the joint action taken by players will constitute a Nash equilibrium. Furthermore, we have shown how to guarantee that a collective objective in a congestion game is a (non-unique) Nash equilibrium.

Our motivation has been that in many engineered systems, the functional forms of utility functions are not available, and so players must adjust their strategies through an adaptive process using only payoff measurements. In the dynamic processes defined here, there is no explicit cooperation or communication between players. One the one hand, this lack of explicit coordination offers an element of robustness to a variety of uncertainties in the strategy adjustment processes. Nonetheless, an interesting future direction would be to investigate to what degree explicit coordination through limited communications could be beneficial.

5.6 Appendix to Chapter 5

5.6.1 Background on Resistance Trees

For a detailed review of the theory of resistance trees, please see [You93]. Let P^0 denote the probability transition matrix for a finite state Markov chain over the state space Z. Consider a "perturbed" process such that the size of the perturbations can be indexed by a scalar $\epsilon > 0$, and let P^ϵ be the associated transition probability matrix. The process P^ϵ is called a *regular perturbed Markov process* if P^ϵ is ergodic for all sufficiently small $\epsilon > 0$ and P^ϵ approaches P^0 at an exponentially smooth rate [You93]. Specifically, the latter condition means that $\forall z, z' \in Z$,

$$\lim_{\epsilon \to 0^+} P^\epsilon_{zz'} = P^0_{zz'},$$

and

$$P^\epsilon_{zz'} > 0 \text{ for some } \epsilon > 0 \;\Rightarrow\; 0 < \lim_{\epsilon \to 0^+} \frac{P^\epsilon_{zz'}}{\epsilon^{r(z \to z')}} < \infty,$$

for some nonnegative real number $r(z \to z')$, which is called the *resistance* of the transition $z \to z'$. (Note in particular that if $P^0_{zz'} > 0$ then $r(z \to z') = 0$.)

Let the recurrence classes of P^0 be denoted by $E_1, E_2, ..., E_N$. For each pair of distinct recurrence classes E_i and E_j, $i \neq j$, an ij-path is defined to be a sequence of distinct states $\zeta = (z_1 \to z_2 \to ... \to z_n)$ such that $z_1 \in E_i$ and $z_n \in E_j$. The resistance of this path is the sum of the resistances of its edges, that is, $r(\zeta) = r(z_1 \to z_2) + r(z_2 \to z_3) + ... + r(z_{n-1} \to z_n)$. Let $\rho_{ij} = \min r(\zeta)$ be the least resistance over all ij-paths ζ. Note that ρ_{ij} must be positive for all distinct i and j, because there exists no path of zero resistance between distinct recurrence classes.

Now construct a complete directed graph with N vertices, one for each recurrence class. The vertex corresponding to class E_j will be called j. The weight on the directed edge $i \to j$ is ρ_{ij}. A tree, T, rooted at vertex j, or j-tree, is a set of $N-1$ directed edges such that, from every vertex different from j, there is a unique directed path in the tree to j. The resistance of a rooted tree, T, is the sum of the resistances ρ_{ij} on the $N-1$ edges that compose it. The *stochastic potential*, γ_j, of the recurrence class E_j is defined to be the minimum resistance over all trees rooted at j. The following theorem gives a simple criterion for determining the stochastically stable states ([You93], Theorem 4).

Theorem 5.6.1. *Let P^ϵ be a regular perturbed Markov process, and for each $\epsilon > 0$ let μ^ϵ be the unique stationary distribution of P^ϵ. Then $\lim_{\epsilon\to 0}\mu^\epsilon$ exists and the limiting distribution μ^0 is a stationary distribution of P^0. The stochastically stable states (i.e., the support of μ^0) are precisely those states contained in the recurrence classes with minimum stochastic potential.*

CHAPTER 6

Connections Between Cooperative Control and Potential Games

In this chapter, we present a view of cooperative control using the language of learning in games. We review the game theoretic concepts of potential games and weakly acyclic games and demonstrate how several cooperative control problems such as consensus and dynamic sensor coverage can be formulated in these settings. Motivated by this connection, we build upon game theoretic concepts to better accommodate a broader class of cooperative control problems. In particular, we extend existing learning algorithms to accommodate restricted action sets caused by limitations in agent capabilities. Furthermore, we also introduce a new class of games, called sometimes weakly acyclic games, for time-varying objective functions and action sets, and provide distributed algorithms for convergence to an equilibrium. Lastly, we illustrate the potential benefits of this connection on several cooperative control problems. For the consensus problem, we demonstrate that consensus can be reached even in an environment with non-convex obstructions. For the functional consensus problem, we demonstrate an approach that will allow agents to reach consensus on a specific consensus point. For the dynamic sensor coverage problem, we demonstrate how autonomous sensors can distribute themselves using only local information in such a way as to maximize the probability of detecting an event over a given mission space. Lastly, we demonstrate how the popular mathematical game of Sudoku can be modeled as a potential game and solved using the learning algorithms discussed in this chapter.

6.1 Introduction

Our goals in this chapter are to establish a relationship between cooperative control problems, such as the consensus problem, and game theoretic methods, and to demonstrate the effectiveness of utilizing game theoretic approaches for controlling multi-agent systems. The results presented here are of independent interest in terms of their applicability to a large class of games. However, we will focus on the consensus problem as the main illustration of the approach.

We consider a discrete time version of the consensus problem initiated in [TBA86] in which a group of players $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_n\}$ seek to come to an agreement, or consensus, upon a common scalar value[1] by repeatedly interacting with one another. By reaching consensus, we mean converging to the agreement space characterized by

$$a_1 = a_2 = \cdots = a_n,$$

where a_i is referred to as the state of player $\mathcal{P}_i$. Several papers study different interaction models and analyze the conditions under whether these interactions lead to consensus [BHO05, XB04, XB05, OM03, OFM07, Mor04, JLM03, KBS06].

A well studied protocol, referred to here as the "consensus algorithm", can be described as follows. At each time step $t \in \{0, 1, \ldots\}$, each player $\mathcal{P}_i$ is allowed to interact with a group of other players, who are referred to as the *neighbors* of player $\mathcal{P}_i$ and denoted as $N_i(t)$. During an interaction, each player $\mathcal{P}_i$ is informed of the current (or possibly delayed) state of all his neighbors. Player $\mathcal{P}_i$ then updates his state by forming a convex combination of his state along with the state of all his neighbors.

[1] The forthcoming results will hold for multi-dimensional consensus as well.

The consensus algorithm takes on the general form

$$a_i(t+1) = \sum_{\mathcal{P}_j \in N_i(t)} \omega_{ij}(t) a_j(t), \tag{6.1}$$

where $\omega_{ij}(t)$ is the relative weight that player $\mathcal{P}_i$ places on the state of player $\mathcal{P}_j$ at time t. The interaction topology is described in terms of a time varying directed graph $G(V, E(t))$ with the set of nodes $V = \mathcal{P}$ and the set of edges $E(t) \subset \mathcal{P} \times \mathcal{P}$ at time t. The set of edges is directly related to the neighbor sets as follows: $(\mathcal{P}_i, \mathcal{P}_j) \in E(t)$ if and only if $\mathcal{P}_j \in N_i(t)$. We will refer to $G(V, E(t))$ as the interaction graph at time t.

There has been extensive research centered on understanding the conditions necessary for guaranteeing the convergence of all states, i.e. $\lim_{t\to\infty} a_i(t) \to a^*$, for all players $\mathcal{P}_i \in \mathcal{P}$. The convergence properties of the consensus algorithm have been studied under several interaction models encompassing delays in information exchange, connectivity issues, varying topologies and noisy measurements.

Surprisingly, there has been relatively little research that links cooperative control problems to a branch of the learning in games literature [You98] that emphasizes coordination games. The goal of this chapter is to better establish this link and to develop new algorithms for broader classes of cooperative control problems as well as games.

In Section 6.2 we establish a connection between cooperative control problems and potential games. In Section 6.3 we model the consensus problem as a potential game and present suitable learning algorithms that guarantee that players will come to a consensus even in an environment filled with non-convex obstructions. In Section 6.4 we introduce a new class of games called sometimes weakly acyclic games, which generalize potential games, and present simple learning dynamics with desirable convergence properties. In Section 6.5 we show that the consensus problem can be modeled as a sometimes weakly acyclic game. In Section 6.6 we develop learning algorithms that can accommodate group based decisions. In Section 6.7 we model the functional consensus problem as a potential game with group based decisions. In Section 6.8 we illustrate the connection between cooperative control and potential games on the dynamic sensor allocation problem and also the mathematical puzzle of Sudoku. Section 6.9 presents some final remarks.

6.2 Cooperative Control Problems and Potential Games

Cooperative control problems entail several autonomous players seeking to collectively accomplish a global objective. The consensus problem is one example of a cooperative control problem, where the global objective is for all players to reach consensus upon a given state. The challenge in cooperative control problems is designing local control laws and/or local objective functions for each of the individual players so that collectively they accomplish the desired global objective.

One approach for cooperative control problems is to assign each individual player a fixed protocol or policy. This protocol specifies precisely what each player should do under any environmental condition. The consensus algorithm set forth in Equation (6.1) is an example of such a policy based approach. One challenge in this approach is to incorporate dynamic or evolving constraints on player policies. For example, suppose a global planner desires a group of autonomous players to physically converge to a central location in an environment containing obstructions. The standard consensus algorithm may not be applicable to this problem since limitations in control capabilities caused by environmental obstructions are not considered. Variations of the consensus algorithm could possibly be designed to accommodate obstructions, but the analysis and control design would be more challenging.

An alternative, *game theoretic* approach to cooperative control problems, and our main interest in this chapter, is to assign each individual player a local objective function. In this setting, each player $\mathcal{P}_i \in \mathcal{P}$ is assigned an action set $\mathcal{A}_i$ and a local objective function $U_i : \mathcal{A} \to \mathbb{R}$, where $\mathcal{A} = \prod_{\mathcal{P}_i \in \mathcal{P}} \mathcal{A}_i$ is the set of joint actions. An example of an objective function that will be studied in the following section is

$$U_i(a_i, a_{-i}) \;\; := \;\; -\sum_{\mathcal{P}_j \in N_i} \|a_i - a_j\|,$$

where $\|\cdot\|$ is any norm, N_i is the neighbor set of player $\mathcal{P}_i$, and $a_{-i} = \{a_1, \ldots, a_{i-i}, a_{i+1}, \ldots, a_n\}$ denotes the collection of actions of players other than player $\mathcal{P}_i$. With this notation, we will frequently express the joint action a as (a_i, a_{-i}).

We are interested in analyzing the long term behavior when players are repeatedly allowed to interact with one another in a competitive environment where each player seeks to selfishly maximize his own objective function. These interactions will be modeled as a repeated game, in which a one stage game is repeated each time step $t \in \{0, 1, 2, \ldots\}$. At every time step $t > 0$, each player $\mathcal{P}_i \in \mathcal{P}$ selects an action $a_i \in \mathcal{A}_i$ seeking to myopically maximize his expected utility. Since a player's utility may be adversely affected by the actions of other players, the player can use his observations from the games played at times $\{0, 1, \ldots, t-1\}$ to develop a behavioral model of the other players.

At any time $t > 0$, the learning dynamics specify how any player $\mathcal{P}_i$ processes past observations from the interactions at times $\{0, 1, \ldots, t-1\}$ to generate a model of the behavior of the other players. The learning dynamics that will be used throughout this chapter are referred to as single stage memory dynamics which have a structural form similar to that of the consensus algorithm; namely, that the decision of any player $\mathcal{P}_i$ at time t is made using only observations from the game played at time $t-1$. The learning dynamics need not be restricted to single stage memory. A follow up study could analyze the benefit of using additional memory in learning dynamics for the consensus problem.

The challenge of the control design for a game theoretic approach lies in designing the objective functions *and* the learning dynamics such that, when players selfishly pursue their own objectives, they also collectively accomplish the objective of the global planner. Suppose that the objective of the global planner is captured by a potential function $\phi : \mathcal{A} \to \mathbb{R}$. In any successful multi-agent system each player's objective function should be appropriately "aligned" with the objective of the global planner. This notion of utility alignment in multi-agent systems has a strong connection to potential games [MS96b]. For convenience, we will restate the definition of potential games originally defined in Section 2.3.2.

Definition 6.2.1 (Potential Games). *Player action sets $\{\mathcal{A}_i\}_{i=1}^n$ together with player objective functions $\{U_i : \mathcal{A} \to \mathbb{R}\}_{i=1}^n$ constitute a* ***potential game*** *if, for some potential function $\phi : \mathcal{A} \to \mathbb{R}$,*

$$U_i(a_i'', a_{-i}) - U_i(a_i', a_{-i}) = \phi(a_i'', a_{-i}) - \phi(a_i', a_{-i}),$$

for every player $\mathcal{P}_i \in \mathcal{P}$, for every $a_i', a_i'' \in \mathcal{A}_i$, and for every $a_{-i} \in \times_{j \neq i} \mathcal{A}_j$.

A potential game, as defined above, requires perfect alignment between the global objective and the players' local objective functions, meaning that if a player unilaterally changed his action, the change in his objective function would be equal to the change in the potential function. There are weaker notions of potential games, called weakly acyclic games, which will be discussed later. The connection between cooperative control problems and potential games is important because learning algorithms for potential games have been studied extensively in the game theory literature [MS96a, MS96b, MS97, MAS07b, MAS05]. Accordingly, if it is shown that a cooperative control problem can be modeled as a potential game, established learning algorithms with guaranteed asymptotic results could be used to tackle the cooperative control problem at hand.

In the following section we will illustrate this opportunity by showing that the consensus problem can be modeled as a potential game by defining players' utilities appropriately.

6.3 Consensus Modeled as a Potential Game

In this section we will formulate the consensus problem as a potential game. First, we establish a global objective function that captures the notion of consensus. Next, we show that local objective functions can be assigned to each player so that the resulting game is in fact a potential game. Finally, we present a learning algorithm that guarantees consensus even in an environment containing non-convex obstructions.

It turns out that the potential game formulation of the consensus problem discussed in this section requires the interaction graph to be time-invariant and undirected. In Section 6.5 we relax these requirements by formulating the consensus problem as a sometimes weakly acyclic game.

6.3.1 Setup: Consensus Problem with a Time-Invariant and Undirected Interaction Graph

Consider a consensus problem with n-player set $\mathcal{P}$ where each player $\mathcal{P}_i \in \mathcal{P}$ has a finite action set $\mathcal{A}_i$. A player's action set could represent the finite set of locations that a player could select.

We will consider the following potential function for the consensus problem

$$\phi(a) := -\sum_{\mathcal{P}_i \in \mathcal{P}} \sum_{\mathcal{P}_j \in N_i} \frac{\|a_i - a_j\|}{2}, \tag{6.2}$$

where $N_i \subset \mathcal{P}$ is player $\mathcal{P}_i$'s time-invariant neighbor set. In the case where the interaction graph induced by the neighbor sets $\{N_i\}_{i=1}^n$ is connected[2], the potential function above achieves the value of 0 if and only if the action profile $a \in \mathcal{A}$ constitutes a *consensus*, i.e.,

$$\phi(a) = 0 \Leftrightarrow a_1 = \cdots = a_n.$$

The goal is to assign each player an objective function that it is perfectly aligned with the global objective in (6.2). One approach would be to assign each player the following objective function:

$$U_i(a) = \phi(a).$$

This assignment would require each player to observe the decision of all players in order to evaluate his payoff for a particular action choice, which may be infeasible. An alternative approach would be to assign each player an objective function that captures the player's marginal contribution to the potential function. For the consensus problem, this translates to each player being assigned the objective function

$$U_i(a_i, a_{-i}) \quad = \quad - \sum_{\mathcal{P}_j \in N_i} \|a_i - a_j\|. \tag{6.3}$$

Now, each player's objective function is *only* dependent on the actions of his neighbors. An objective function of this form is referred to as Wonderful Life Utility; see [AMS07, WT99]. It is known that assigning each agent a Wonderful Life Utility leads to a potential game [AMS07, WT99]; however, we will explicitly show this for the consensus problem in the following claim.

Claim 6.3.1. *Player objective functions (6.3) constitute a potential game with the potential function (6.2) provided that the time-invariant interaction graph induced by the neighbor sets $\{N_i\}_{i=1}^n$ is undirected, i.e.,*

$$\mathcal{P}_j \in N_i \Leftrightarrow \mathcal{P}_i \in N_j.$$

[2] A graph is connected if there exists a path from any node to any other node.

Proof. Since the interaction graph is time-invariant and undirected, the potential function can be expressed as

$$\phi(a) = -\sum_{\mathcal{P}_j \in N_i} \|a_i - a_j\| - \sum_{\mathcal{P}_j \neq \mathcal{P}_i} \sum_{\mathcal{P}_k \in N_j \backslash \mathcal{P}_i} \frac{\|a_j - a_k\|}{2}.$$

The change in objective of player $\mathcal{P}_i$ by switching from action a_i^1 to action a_i^2 provided that all other players collectively play a_{-i} is

$$\begin{aligned} U_i(a_i^2, a_{-i}) - U_i(a_i^1, a_{-i}) &= \sum_{\mathcal{P}_j \in N_i} -\|a_i^2 - a_j\| + \|a_i^1 - a_j\|, \\ &= \phi(a_i^2, a_{-i}) - \phi(a_i^1, a_{-i}). \end{aligned}$$

□

Note that the above claim does not require the interaction graph to be connected. There may exist other potential functions and subsequent player objective functions that can accommodate more general setups. For a detailed discussion on possible player objective functions derived from a given potential function, see [AMS07].

We now assume that the above game is repeatedly played at discrete time steps $t \in \{0, 1, 2, \ldots\}$. We are interested in determining the limiting behavior of the players, in particular whether or not they reach a consensus, under various interaction models. Since the consensus problem is modeled as a potential game, there are a large number of learning algorithms available with guaranteed results [You98, You05, AMS07, MS96b, MAS07b, MAS05]. Most of the learning algorithms for potential games guarantee that the player behavior converges to a Nash equilibrium.

It is straightforward to see that any consensus point is a Nash equilibrium of the game characterized by the player objective functions (6.3). This is because a consensus point maximizes the potential function as well as the player objective functions (6.3). However, the converse statement is not true. Let $\mathcal{A}^*$ denote the set of Nash equilibria and $\mathcal{A}^c$ denote the set of consensus points. We know that $\mathcal{A}^c \subset \mathcal{A}^*$ where the inclusion can be proper. In other words, a Nash equilibrium, say $a^* \in \mathcal{A}^*$, can be suboptimal, i.e., $\phi(a^*) < 0$, and hence fail to be a consensus point.

6.3.2 A Learning Algorithm for Potential Games with Suboptimal Nash Equilibria

Before stating the learning algorithm, we start with some notation. Let the strategy of player $\mathcal{P}_i$ at time t be denoted by the probability distribution $p_i(t) \in \Delta(\mathcal{A}_i)$ where $\Delta(\mathcal{A}_i)$ denotes the set of probability distributions over the set $\mathcal{A}_i$. Using this strategy, player $\mathcal{P}_i$ randomly selects an action from $\mathcal{A}_i$ at time t according to $p_i(t)$.

Consider the following learning algorithm known as spatial adaptive play (SAP) [You98]. At each time $t > 0$, one player $\mathcal{P}_i \in \mathcal{P}$ is randomly chosen (with equal probability for each player) and allowed to update his action. All other players must repeat their actions, i.e. $a_{-i}(t) = a_{-i}(t-1)$. At time t, the updating player $\mathcal{P}_i$ randomly selects an action from $\mathcal{A}_i$ according to his strategy $p_i(t) \in \Delta(\mathcal{A}_i)$ where the a_i-th component $p_i^{a_i}(t)$ of his strategy is given as

$$p_i^{a_i}(t) = \frac{\exp\{\beta\, U_i(a_i, a_{-i}(t-1))\}}{\sum_{\bar{a}_i \in \mathcal{A}_i} \exp\{\beta\, U_i(\bar{a}_i, a_{-i}(t-1))\}},$$

for some exploration parameter $\beta \geq 0$. The constant β determines how likely player $\mathcal{P}_i$ is to select a suboptimal action. If $\beta = 0$, player $\mathcal{P}_i$ will select any action $a_i \in \mathcal{A}_i$ with equal probability. As $\beta \to \infty$, player $\mathcal{P}_i$ will select an action from his best response set

$$\{a_i \in \mathcal{A}_i : U_i(a_i, a_{-i}(t-1)) = \max_{a_i' \in \mathcal{A}_i} U_i(a_i', a_{-i}(t-1))\}$$

with arbitrarily high probability.

In a repeated potential game in which all players adhere to SAP, the stationary distribution $\mu \in \Delta(\mathcal{A})$ of the joint action profiles is given in [You98] as

$$\mu(a) = \frac{\exp\{\beta\ \phi(a)\}}{\sum_{\bar{a}\in\mathcal{A}} \exp\{\beta\ \phi(\bar{a})\}}.$$

One can interpret the stationary distribution μ as follows: for sufficiently large times $t > 0,\ \mu(a)$ equals the probability that $a(t) = a$. As $\beta \uparrow \infty$, all the weight of the stationary distribution μ is on the joint actions that maximize the potential function. In the potential game formulation of the consensus problem, the joint actions that maximize the potential function (6.2) are precisely the consensus points provided that the interaction graph is connected. Therefore, if all players update their actions using the learning algorithm SAP with sufficiently large β, then the players will reach a consensus asymptotically with arbitrarily high probability.

6.3.3 A Learning Algorithm for Potential Games with Suboptimal Nash Equilibria and Restricted Action Sets

One issue with the applicability of the learning algorithm SAP for the consensus problem is that it permits any player to select any action in his action set. Because of player mobility limitations, this may not be possible. For example, a player may only be able to move to a position within a fixed radius of his current position. Therefore, we seek to modify SAP by conditioning a player's action set on his previous action. Let $a(t-1)$ be the joint action at time $t-1$. With restricted action sets, the set of actions available to player $\mathcal{P}_i$ at time t is a function of his action at time $t-1$ and will be denoted as $R_i(a_i(t-1)) \subset \mathcal{A}_i$. We will adopt the convention that $a_i \in R_i(a_i)$ for any action $a_i \in \mathcal{A}_i$, i.e., a player is always allowed to stay with his previous action.

We will introduce a variant of SAP called binary Restrictive Spatial Adaptive Play (RSAP) to accommodate the notion of restricted action sets. RSAP can be described as follows: At each time step $t > 0$, one player $\mathcal{P}_i \in \mathcal{P}$ is randomly chosen (with equal probability for each player) and allowed to update his action. All other players must repeat their actions, i.e. $a_{-i}(t) = a_{-i}(t-1)$. At time t, the updating player $\mathcal{P}_i$ selects one trial action $\hat{a}_i$ randomly from his allowable set $R_i(a_i(t-1))$ with the following probability:

- $\mathbf{Pr}\left[\hat{a}_i = a_i\right] = \frac{1}{N_i}$ for any $a_i \in R_i(a_i(t-1)) \setminus a_i(t-1)$,
- $\mathbf{Pr}\left[\hat{a}_i = a_i(t-1)\right] = 1 - \frac{|R_i(a_i(t-1))|-1}{N_i}$,

where N_i denotes the maximum number of actions in any restricted action set for player $\mathcal{P}_i$, i.e., $N_i := \max_{a_i \in \mathcal{A}_i} |R_i(a_i)|$. After player $\mathcal{P}_i$ selects a trial action $\hat{a}_i$, he chooses his action at time t as follows:

$$\begin{aligned}
\mathbf{Pr}\left[a_i(t) = \hat{a}_i\right] &= \frac{\exp\{\beta\, U_i(\hat{a}_i, a_{-i}(t-1))\}}{\exp\{\beta\, U_i(\hat{a}_i, a_{-i}(t-1))\} + \exp\{\beta\, U_i(a(t-1))\}}, \\
\mathbf{Pr}\left[a_i(t) = a_i(t-1)\right] &= \frac{\exp\{\beta\, U_i(a(t-1))\}}{\exp\{\beta\, U_i(\hat{a}_i, a_{-i}(t-1))\} + \exp\{\beta\, U_i(a(t-1))\}},
\end{aligned}$$

where $\beta \geq 0$ is an exploration parameter. Note that if $\hat{a}_i$ is selected as $a_i(t-1)$ then $\mathbf{Pr}\left[a_i(t) = a_i(t-1)\right] = 1$.

We make the following assumptions regarding the restricted action sets.

Assumption 6.3.1 (Reversibility). *For any player $\mathcal{P}_i \in \mathcal{P}$ and any action pair $a_i^1, a_i^2 \in \mathcal{A}_i$,*

$$a_i^2 \in R_i(a_i^1) \Leftrightarrow a_i^1 \in R_i(a_i^2).$$

Assumption 6.3.2 (Feasibility). *For any player $\mathcal{P}_i \in \mathcal{P}$ and any action pair $a_i^0, a_i^n \in \mathcal{A}_i$, there exists a sequence of actions $a_i^0 \rightarrow a_i^1 \rightarrow \cdots \rightarrow a_i^n$ satisfying $a_i^k \in R_i(a_i^{k-1})$ for all $k \in \{1, 2, \ldots, n\}$.*

Theorem 6.3.1. *Consider a finite n-player potential game with potential function $\phi(\cdot)$. If the restricted action sets satisfy Assumptions 6.3.1 and 6.3.2, then RSAP induces* *a Markov process over the state space $\mathcal{A}$ where the unique stationary distribution $\mu \in \Delta(\mathcal{A})$ is given as*

$$\mu(a) = \frac{\exp\{\beta\ \phi(a)\}}{\sum_{\bar{a}\in\mathcal{A}} \exp\{\beta\ \phi(\bar{a})\}},\ \textit{for any } a \in \mathcal{A}. \tag{6.4}$$

Proof. The proof follows along the lines of the proof of Theorem 6.2 in [You98]. By Assumptions 6.3.1 and 6.3.2 we know that the Markov process induced by RSAP is irreducible and aperiodic; therefore, the process has a unique stationary distribution. Below, we show that this unique distribution must be (6.4) by verifying that the distribution (6.4) satisfies the detailed balanced equations

$$\mu(a)P_{ab} = \mu(b)P_{ba},$$

for any $a, b \in \mathcal{A}$, where

$$P_{ab} := \mathbf{Pr}\left[a(t) = b | a(t-1) = a\right].$$

Note that the only nontrivial case is the one where a and b differ by exactly one player $\mathcal{P}_i$, that is, $a_{-i} = b_{-i}$ but $a_i \neq b_i$ where $a_i \in R_i(b_i)$ which also implies that $b_i \in R_i(a_i)$. Since player $\mathcal{P}_i$ has probability $1/n$ of being chosen in any given period and any trial action $b_i \in R_i(a_i)$, $b_i \neq a_i$, has probability of $1/N_i$ of being chosen, it follows that

$$\mu(a)P_{ab} = \left[\frac{\exp\{\beta\ \phi(a)\}}{\sum_{z\in\mathcal{A}} \exp\{\beta\ \phi(z)\}}\right] \times \left[(1/n)(1/N_i)\frac{\exp\{\beta\ U_i(b)\}}{\exp\{\beta\ U_i(a)\} + \exp\{\beta\ U_i(b)\}}\right].$$

Letting

$$\lambda = \left(\frac{1}{\sum_{z\in\mathcal{A}} \exp\{\beta\ \phi(z)\}}\right) \times \left(\frac{(1/n)(1/N_i)}{\exp\{\beta\ U_i(a)\} + \exp\{\beta\ U_i(b)\}}\right),$$

we obtain

$$\mu(a)P_{ab} = \lambda \exp\{\beta\phi(a) + \beta U_i(b)\}.$$

Since $U_i(b) - U_i(a) = \phi(b) - \phi(a)$, we have

$$\mu(a)P_{ab} = \lambda \exp\{\beta\phi(b) + \beta U_i(a)\},$$

which leads us to

$$\mu(a)P_{ab} = \mu(b)P_{ba}.$$

□

Note that if all players adhere to the learning dynamics RSAP in a consensus problem where the interaction graph is time-invariant and undirected, the restricted action sets satisfy Assumptions 6.3.1 and 6.3.2, and players are assigned the utilities (6.3), then, at sufficiently large times t, the players' collective behavior will maximize the potential function (6.2) with arbitrarily high probability provided that β is sufficiently large. Furthermore, if the interaction graph is connected and consensus is possible, meaning $(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \cdots \cap \mathcal{A}_n) \neq \emptyset$, then, at sufficiently large times $t > 0$, the players' actions will constitute a consensus with arbitrarily high probability even in an environment filled with non-convex obstructions.

6.3.4 Example: Consensus in an Environment with Non-convex Obstructions

Consider the 2-D consensus problem with player set $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4\}$. Each player $\mathcal{P}_i$ has an action set $\mathcal{A}_i = \{1, 2, \ldots, 10\} \times \{1, 2, .., 10\}$ as illustrated in Figure 6.1. The arrows represent the time-invariant and undirected edges of the connected interaction graph. The restricted action sets are highlighted for players $\mathcal{P}_2$ and $\mathcal{P}_4$. At any given time, any player can have at most 9 possible actions; therefore, $N_i = 9$ for all players $\mathcal{P}_i \in \mathcal{P}$.

We simulated RSAP on the consensus problem with the interaction graph, environmental obstruction, and the initial conditions shown in Figure 6.1. We increase the exploration parameter β as $t/200$ during player interactions. The complete action path of all players reaching a consensus is shown in Figure 6.2.

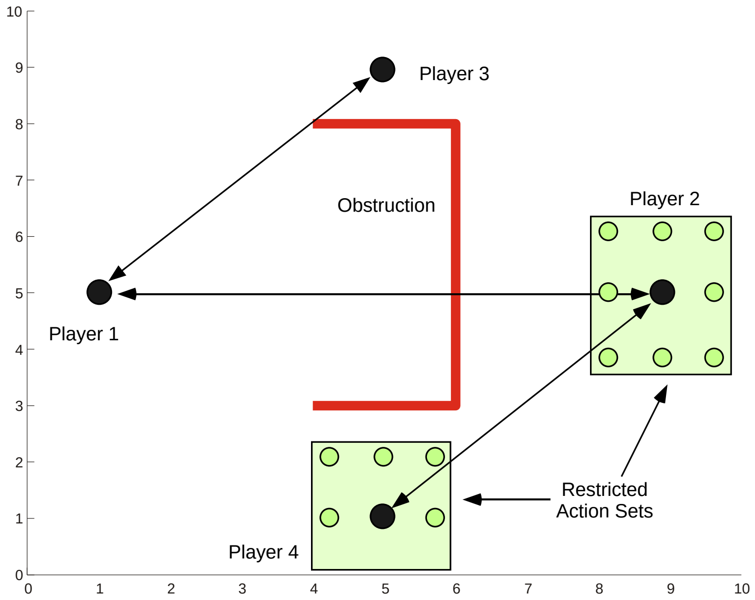


Figure 6.1: Example: Setup of a Consensus Problem with Restricted Action Sets and Non-convex Environmental Obstructions.

6.4 Weakly Acyclic and Sometimes Weakly Acyclic Games

In potential games, players' objective functions must be perfectly aligned with the potential of the game. In the potential game formulation of the consensus problem, this alignment condition required that the interaction graph be time-invariant and undirected. In this section we will seek to relax this alignment requirement by allowing player objective functions to be "somewhat" aligned with the potential of the game. We will review a weaker form of potential games called weakly acyclic games and introduce a new class of games called *sometimes weakly acyclic games*. We will also present simple learning dynamics that guarantee convergence to a *universal Nash equilibrium*, to be defined later, in any sometimes weakly acyclic game.

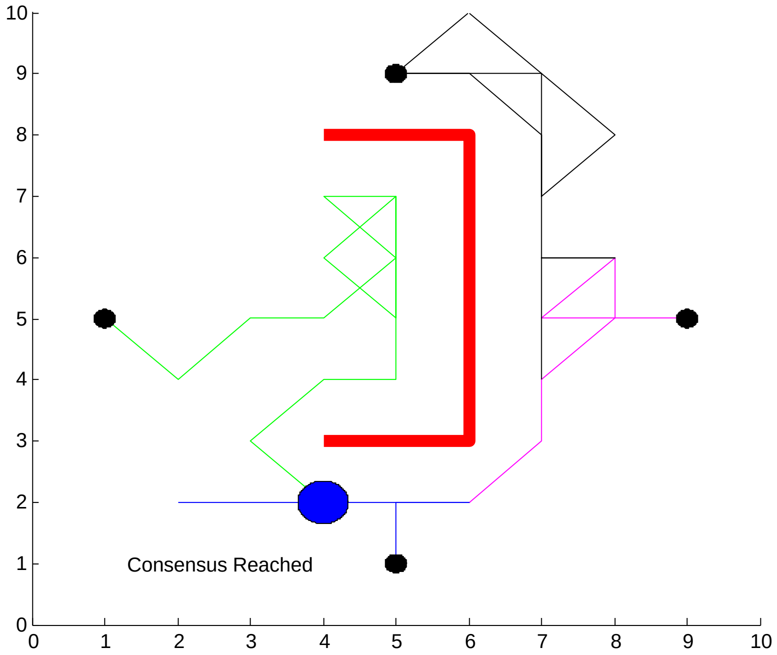


Figure 6.2: Example: Evolution of the Action Path in the Consensus Problem with Restricted Action Sets and Non-convex Environmental Obstructions.

6.4.1 Weakly Acyclic Games

Recall the definition of a weakly acyclic game from Section 2.3.4. A game is weakly acyclic if, for any $a \in \mathcal{A}$, there exists a better reply path starting at a and ending at some Nash equilibrium [You98, You05].

The above definition does not clearly identify the similarities between potential games and weakly acyclic games. Furthermore, using this definition to show that a given game G, i.e., the players, objective functions, and action sets, is weakly acyclic would be problematic. With these issues in mind, we will now derive an equivalent definition for weakly acyclic games that utilizes potential functions.

Lemma 6.4.1. *A game is weakly acyclic if and only if there exists a potential function* $\phi : \mathcal{A} \rightarrow \mathbb{R}$ *such that for any action* $a \in \mathcal{A}$ *that is not a Nash equilibrium, there* *exists a player $\mathcal{P}_i \in \mathcal{P}$ with an action $a_i^* \in \mathcal{A}_i$ such that $U_i(a_i^*, a_{-i}) > U_i(a_i, a_{-i})$ and $\phi(a_i^*, a_{-i}) > \phi(a_i, a_{-i})$.*

Proof. ($\Leftarrow$) Select any action $a^0 \in \mathcal{A}$. If a^0 is not a Nash equilibrium, there exists a player $\mathcal{P}_i \in \mathcal{P}$ with an action $a_i^* \in \mathcal{A}_i$ such that $U_i(a^1) > U_i(a^0)$ and $\phi(a^1) > \phi(a^0)$ where $a^1 = (a_i^*, a_{-i}^0)$.

Repeat this process and construct a path $a^0, a^1, \ldots, a^n$ until it can no longer be extended. Note first that such a path cannot cycle back on itself, because ϕ is strictly increasing along the path. Since $\mathcal{A}$ is finite, the path cannot be extended indefinitely. Hence, the last element in this path must be a Nash equilibrium.

($\Rightarrow$) We will construct a potential function $\phi : \mathcal{A} \to \mathbb{R}$ recursively. Select any action $a^0 \in \mathcal{A}$. Since the game is weakly acyclic, there exists a better reply path $a^0, a^1, \ldots, a^n$ where a^n is a Nash equilibrium. Let $\mathcal{A}^0 = \{a^0, a^1, \ldots, a^n\}$. Define the (finite) potential function ϕ over the set $\mathcal{A}^0$ satisfying the following conditions:

$$\phi(a^0) < \phi(a^1) < \cdots < \phi(a^n).$$

Now select any action $\tilde{a}^0 \in \mathcal{A} \setminus \mathcal{A}^0$. There exists a better reply path $\tilde{a}^0, \tilde{a}^1, \ldots, \tilde{a}^m$ where $\tilde{a}^m$ is a Nash equilibrium. Let $\mathcal{A}^1 = \{\tilde{a}^0, \tilde{a}^1, \ldots, \tilde{a}^m\}$. If $\mathcal{A}^1 \cap \mathcal{A}^0 = \emptyset$ then define the potential function ϕ over the set $\mathcal{A}^1$ satisfying the following conditions:

$$\phi(\tilde{a}^0) < \phi(\tilde{a}^1) < \cdots < \phi(\tilde{a}^m).$$

If $\mathcal{A}^1 \cap \mathcal{A}^0 \neq \emptyset$, then let $k^* = \min\{k \in \{1, 2, \ldots, m\} : \tilde{a}^k \in \mathcal{A}^0\}$. Define the potential function ϕ over the truncated (redefined) set $\mathcal{A}^1 = \{\tilde{a}^0, \tilde{a}^1, \ldots, \tilde{a}^{k^*-1}\}$ satisfying the following conditions:

$$\phi(\tilde{a}^0) < \phi(\tilde{a}^1) < \cdots < \phi(\tilde{a}^{k^*}).$$

Now select any action $\hat{a}^0 \in \mathcal{A} \setminus (\mathcal{A}^0 \cup \mathcal{A}^1)$ and repeat until no such action exists.

The construction of the potential function ϕ guarantees that for any action $a \in \mathcal{A}$ that is not a Nash equilibrium, there exists a player $\mathcal{P}_i \in \mathcal{P}$ with an action $a_i^* \in \mathcal{A}_i$ such that $U_i(a_i^*, a_{-i}) > U_i(a_i, a_{-i})$ and $\phi(a_i^*, a_{-i}) > \phi(a_i, a_{-i})$. □

6.4.2 Learning Dynamics for Weakly Acyclic Games

We will consider the *better reply with inertia dynamics* for weakly acyclic games analyzed in [You93, You05]. Before stating the learning dynamics, we define a player's *strict better reply set* for any action profile $a^0 \in \mathcal{A}$ as

$$B_i(a^0) := \{a_i \in \mathcal{A}_i : U_i(a_i, a_{-i}^0) > U_i(a^0)\}.$$

The better reply with inertia dynamics can be described as follows. At each time $t > 0$, each player $\mathcal{P}_i$ presumes that all other players will continue to play their previous actions $a_{-i}(t-1)$. Under this presumption, each player $\mathcal{P}_i \in \mathcal{P}$ selects an action according to the following strategy at time t:

$$\begin{aligned} B_i(a(t-1)) = \emptyset \quad &\Rightarrow \quad a_i(t) = a_i(t-1), \\ B_i(a(t-1)) \neq \emptyset \quad &\Rightarrow \quad \begin{cases} \mathbf{Pr}\left[a_i(t) = a_i(t-1)\right] = \alpha(t), \\ \mathbf{Pr}\left[a_i(t) = a_i^*\right] = \frac{1-\alpha(t)}{|B_i(a(t-1))|}, \end{cases} \end{aligned}$$

for any action $a_i^* \in B_i(a(t-1))$ where $\alpha(t) \in (0,1)$ is referred to as the player's inertia at time t. According to these rules, player $\mathcal{P}_i$ will stay with the previous action $a_i(t-1)$ with probability $\alpha(t)$ even when there is a perceived opportunity for improvement. We make the following standing assumption on the players' willingness to optimize.

Assumption 6.4.1. *There exist constants $\underline{\varepsilon}$ and $\bar{\varepsilon}$ such that for all time $t \geq 0$ and for all players $\mathcal{P}_i \in \mathcal{P}$,*

$$0 < \underline{\varepsilon} < \alpha_i(t) < \bar{\varepsilon} < 1.$$

This assumption implies that players are always willing to optimize with some nonzero inertia.

If all players adhere to the better reply with inertia dynamics satisfying Assumption 6.4.1, then the joint action profiles will converge to a Nash equilibrium almost surely in any weakly acyclic game [You93, You05].

6.4.3 Sometimes Weakly Acyclic Games

In the potential game formulation of the consensus problem, each player was assigned a time-invariant objective function of the form (6.3). However, in the case of a time-varying interaction topology, we would like to allow player objective functions to be time-varying. In this framework, each player $\mathcal{P}_i$ is now assigned a local objective function $U_i : \mathcal{A} \times \{0, 1, 2, \dots\} \to \mathbb{R}$. We will denote the objective function of player $\mathcal{P}_i$ at time t as $U_i(a(t), t)$ where $a(t)$ is the action profile at time t.

We will call an action profile a^* a *universal Nash equilibrium* if

$$U_i(a^*, t) = \max_{a_i \in \mathcal{A}_i} U_i(a_i, a^*_{-i}, t)$$

for all times $t \geq 0$.

We will call a game *sometimes weakly acyclic* if there exists a potential function $\phi : \mathcal{A} \to \mathbb{R}$ and a finite time constant T such that for any time $t_0 > 0$ and any action profile a^0 that is *not* a universal Nash equilibrium, there exists a time $t_1 \in [t_0, t_0 + T]$, a player $\mathcal{P}_i \in \mathcal{P}$, and an action $a_i^* \in \mathcal{A}_i$ such that $U_i(a_i^*, a^0_{-i}, t_1) > U_i(a^0, t_1)$ and $\phi(a_i^*, a^0_{-i}) > \phi(a^0)$.

Note that a sometimes weakly acyclic game has at least one universal Nash equilibrium: namely, an action profile that maximizes the potential function phi.

6.4.4 Learning Dynamics for Sometimes Weakly Acyclic Games

We will consider the better reply with inertia dynamics for games involving time-varying objective functions. Before stating the learning dynamics, we redefine a player's *strict better reply set* for any action profile $a^0 \in \mathcal{A}$ and time $t > 0$ as

$$B_i(a^0, t) := \{a_i \in \mathcal{A}_i : U_i(a_i, a^0_{-i}, t) > U_i(a^0, t)\}.$$

The better reply with inertia dynamics can be described as follows. At each time $t > 0$, each player $\mathcal{P}_i$ presumes that all other players will continue to play their previous actions $a_{-i}(t-1)$. Under this presumption, each player $\mathcal{P}_i \in \mathcal{P}$ selects an action according to the following strategy at time t:

$$\begin{aligned} B_i(a(t-1), t) = \emptyset \ &\Rightarrow \ a_i(t) = a_i(t-1), \\ B_i(a(t-1), t) \neq \emptyset \ &\Rightarrow \ \begin{cases} \mathbf{Pr}\left[a_i(t) = a_i(t-1)\right] = \alpha(t), \\ \mathbf{Pr}\left[a_i(t) = a_i^*\right] = \frac{(1-\alpha(t))}{|B_i(a(t-1),t)|}, \end{cases} \end{aligned}$$

for any action $a_i^* \in B_i(a(t-1), t)$ where $\alpha(t) \in (0,1)$ is the player's inertia at time t.

Theorem 6.4.1. *Consider an n-player sometimes weakly acyclic game with finite action sets. If all players adhere to the better reply with inertia dynamics satisfying Assumption 6.4.1, then the joint action profiles will converge to a universal Nash equilibrium almost surely.*

Proof. Let $\phi : \mathcal{A} \to \mathbb{R}$ and T be the potential function and time constant for the sometimes weakly acyclic game. Let $a(t_0) = a^0$ be the action profile at time t_0. If a^0 is a universal Nash equilibrium, then $a(t) = a^0$ for all times $t \geq t_0$ and we are done. Otherwise, there exists a time t_1 satisfying $(t_0 + T) \geq t_1 > t_0$, a player $\mathcal{P}_i \in \mathcal{P}$, and an action $a_i^* \in \mathcal{A}_i$ such that $U_i(a_i^*, a^0_{-i}, t_1) > U_i(a^0, t_1)$ and $\phi(a_i^*, a^0_{-i}) > \phi(a^0)$. Because of players' inertia, the action $a^1 = (a_i^*, a^0_{-i})$ will be played at time t_1 with at least probability $\underline{\epsilon}^{n-1}\frac{(1-\bar{\epsilon})}{|\mathcal{A}|}\underline{\epsilon}^{nT}$.

One can repeat this argument to show that for any time $t_0 > 0$ and any action profile $a(t_0)$ there exists a joint action a^* such that

$$\mathbf{Pr}\left[a(t) = a^*, \forall t \geq t^*\right] \geq \epsilon^*$$

where

$$\begin{aligned} t^* &= t_0 + T\,|\mathcal{A}|, \\ \epsilon^* &= \left(\underline{\epsilon}^{n-1}\frac{(1-\bar{\epsilon})}{|\mathcal{A}|}\underline{\epsilon}^{n\,T}\right)^{|\mathcal{A}|}. \end{aligned}$$

□

6.5 Consensus Modeled as a Sometimes Weakly Acyclic Game

Two main problems arose in the potential game formulation of the consensus problem. The first problem was that a Nash equilibrium was not necessarily a consensus point even when the interaction graph was connected and the environment was obstruction free. Therefore, we needed to employ a stochastic learning algorithm like SAP or RSAP to guarantee that the collective behavior of the players would be a consensus point with arbitrarily high probability. SAP or RSAP led to consensus by introducing noise into the decision making process, meaning that a player would occasionally make a suboptimal choice. The second problem was that the interaction graph needed to be time-invariant, undirected, and connected in order to guarantee consensus.

In this section, we will illustrate that by modeling the consensus problem as a sometimes weakly acyclic game one can effectively alleviate both problems. For brevity, we will show that the 1-dimensional consensus problem with appropriately designed player objective functions is a sometimes weakly acyclic game. However, one can easily extend this to the multi-dimensional case.

6.5.1 Setup: Consensus Problem with a Time-Varying and Directed Interaction Graph

Consider a consensus problem with a n-player set $\mathcal{P}$ and a time-varying and directed interaction graph. Each player has a finite action set $\mathcal{A}_i \subset \mathbb{R}$ and without loss of generalities, we will assume that $\mathcal{A}_1 = \mathcal{A}_2 = \cdots = \mathcal{A}_n$. Each player $\mathcal{P}_i \in \mathcal{P}$ is assigned an objective function $U_i : \mathcal{A} \times \{0, 1, 2, ...\} \to \mathbb{R}$. We make the following standard assumption on players' neighbor sets.

Assumption 6.5.1. *Players' neighbor sets satisfy*

$$\mathcal{P}_i \in N_i(t),\ \ \forall \mathcal{P}_i \in \mathcal{P},\ t > 0.$$

Before introducing the player objective functions, we define the following measure

$$D(a, \mathcal{P}') \ \ := \ \ \max_{\mathcal{P}_i, \mathcal{P}_j \in \mathcal{P}'} (a_i - a_j), \tag{6.5}$$

where $\mathcal{P}' \subseteq \mathcal{P}$, and extreme player sets

$$\begin{aligned} \mathcal{P}^u(a) &:= \{\mathcal{P}_i \in \mathcal{P} : a_i = \max_{\mathcal{P}_j \in \mathcal{P}} a_j\}, \\ \mathcal{P}^l(a) &:= \{\mathcal{P}_i \in \mathcal{P} : a_i = \min_{\mathcal{P}_j \in \mathcal{P}} a_j\}, \\ \underline{n}(a) &:= \min\{|\mathcal{P}^u(a)|, |\mathcal{P}^l(a)|\}. \end{aligned}$$

We also define the constant $\delta_A > 0$ as follows. For any $a^1, a^2 \in \mathcal{A}$ and any player sets $\mathcal{P}^1, \mathcal{P}^2 \subset \mathcal{P}$ such that $D(a^1, \mathcal{P}^1) \neq D(a^2, \mathcal{P}^2)$, the following inequality is satisfied:

$$|D(a^1, \mathcal{P}^1) - D(a^2, \mathcal{P}^2)| > \delta_A.$$

Consider the following potential function $\phi : \mathcal{A} \to \mathbb{R}$

$$\phi(a) = -D(a, \mathcal{P}) + \delta_A(1 - \underline{n}(a)/n). \tag{6.6}$$

Note that the potential function is a non-positive function that achieves the value of 0 if and only if the action profile constitutes a consensus. Furthermore, note that the potential function is independent of the interaction topology.

Rather than specifying a particular objective functions as in (6.3), we will introduce a class of admissible objective functions. To that end, we define the set of *reasonable actions* for player $\mathcal{P}_i$ at time t given any action profile $a^0 \in \mathcal{A}$ as

$$S_i(a^0, t) := \{a_i \in \mathcal{A}_i : \max_{\mathcal{P}_j \in N_i(t)} a_j^0 \geq a_i \geq \min_{\mathcal{P}_k \in N_i(t)} a_k^0\}.$$

Note that

$$a_i \in S_i(a^0, t) \;\Rightarrow\; D(a_i, a_{-i}^0, N_i(t)) \leq D(a^0, N_i(t)).$$

We will define a general class of *reasonable objective functions*. An objective function for player $\mathcal{P}_i$ is called a reasonable objective function if, for any time $t > 0$, and any action profile $a \in \mathcal{A}$, the better response set satisfies the following two conditions:

1. $B_i(a, t) \subset \{S_i(a, t), \emptyset\}$,
2. $|S_i(a, t)| > 1 \;\Rightarrow\; B_i(a, t) \neq \emptyset$.

Roughly speaking, these conditions ensure that a player will not value moving further away from his belief about the location of his neighbors.

We will now relax our requirements on the connectivity and time-invariance of the interaction graph in the consensus problem. A common assumption on the interaction graph is connectedness over intervals.

Assumption 6.5.2 (Connectedness Over Intervals). *There exists a constant $T > 0$ such that for any time $t > 0$, the interaction graph with nodes $\mathcal{P}$ and edges $E = E(t) \cup \cdots \cup E(t + T)$ is connected.*

Claim 6.5.1. *Reasonable objective functions introduced above constitute a sometimes weakly acyclic game with the potential function (6.6) provided that the interaction* *graph satisfies Assumption 6.5.2. Furthermore, every universal Nash equilibrium constitutes consensus.*

Proof. It is easy to see that any consensus point is a universal Nash equilibrium. We will show that if an action profile is not a consensus point, then there exists a player who can increase his objective function as well as the potential function at some time in a fixed time window. This implies that every universal Nash equilibrium is a consensus point and furthermore that the game is sometime weakly acyclic.

Let $a^0 = a(t_0)$ be any joint action that is not a consensus point. We will show that for some time $t_1 \in [t_0, t_0 + T]$ there exists a player $\mathcal{P}_i \in \mathcal{P}$ with an action $a_i^* \in \mathcal{A}_i$ such that $U_i(a_i^*, a_{-i}^0, t_1) > U_i(a^0, t_1)$ and $\phi(a_i^*, a_{-i}^0) > \phi(a^0)$. To see this let $\mathcal{P}^*$ be the extreme player set with the least number of players, i.e., $\mathcal{P}^* = \mathcal{P}^u(a^0)$ if $|\mathcal{P}^u(a^0)| \leq |\mathcal{P}^l(a^0)|$ or $\mathcal{P}^* = \mathcal{P}^l(a^0)$ if $|\mathcal{P}^u(a^0)| > |\mathcal{P}^l(a^0)|$. Since the interaction graph satisfies Assumption 6.5.2[3], for some $t_1 \in [t_0, t_0 + T]$ there exists at least one player $\mathcal{P}_i \in \mathcal{P}^*$ with a neighbor $\mathcal{P}_j \in N_i(t_1)$ such that $\mathcal{P}_j \notin \mathcal{P}^*$. Therefore,

$$|S_i(a^0, t_1)| > 1 \ \Rightarrow |B_i(a^0, t_1)| \neq \emptyset.$$

Let $a_i^* \in B_i(a^0, t_1)$ and for notional convenience let $a^1 = (a_i^*, a_{-i}^0)$. We know that $D(a^1, \mathcal{P}) \leq D(a^0, \mathcal{P})$. If $D(a^1, \mathcal{P}) < D(a^0, \mathcal{P})$, then

$$\begin{aligned}
\phi(a^1) &= -D(a^1, \mathcal{P}) + \delta_A(1 - \underline{n}(a^1)/n), \\
&> -D(a^0, \mathcal{P}) + \delta_A + \delta_A(1 - \underline{n}(a^1)/n), \\
&> -D(a^0, \mathcal{P}) + \delta_A + \delta_A(1 - (\underline{n}(a^0) + n)/n), \\
&= \phi(a^0).
\end{aligned}$$

[3] Note that assumption 6.5.2 is stronger than necessary for this proof.

If $D(a^1, \mathcal{P}) = D(a^0, \mathcal{P})$, then

$$\begin{array}{rcl} \phi(a^1) & = & -D(a^0, \mathcal{P}) + \delta_A(1 - \underline{n}(a^1)/n), \\ & > & -D(a^0, \mathcal{P}) + \delta_A(1 - (\underline{n}(a^1) + 1)/n), \\ & \geq & -D(a^0, \mathcal{P}) + \delta_A(1 - \underline{n}(a^0)/n), \\ & = & \phi(a^0). \end{array}$$

Therefore, a^0 is not a universal Nash equilibrium. □

If all players adhere to the better reply with inertia dynamics in a consensus problem where the interaction graph satisfies Assumption 6.5.2 and the players are assigned reasonable objective functions then the joint action profile will converge almost surely to a consensus point.

These results can easily be extended to a multi-dimensional consensus problem with bounded observational delays.

6.5.2 Extension to Multi-Dimensional Consensus

One can easily extend the arguments above to show that any k-dimensional consensus game is a sometimes weakly acyclic game by generalizing the measure and choosing the extreme player sets appropriately. An example of an acceptable measure is

$$D(a, \mathcal{P}') := \sum_{k=1}^{n} \max_{\mathcal{P}_i, \mathcal{P}_j \in \mathcal{P}'} d_k^T (a_i - a_j).$$

where $\mathcal{P}' \subseteq \mathcal{P}$ and $d_1, d_2, ..., d_n \in \mathbb{R}^k$ is a set of measure vectors which span the complete space of $\mathbb{R}^k$. The term $\max_{\mathcal{P}_i, \mathcal{P}_j \in \mathcal{P}'} d_k^T (a_i - a_j)$ captures the maximum distance between the action of any two agents in the nonempty player set $\mathcal{P}'$ relative to the measure direction d_k. In the 1-D consensus problem, where $d_1 = 1$, the measure reverts back to (6.5).

The set of reasonable actions for player $\mathcal{P}_i$ at time t given the joint action profile a is now

$$S_i(a,t) = \{a_i' \in \mathcal{A}_i : \forall k, \max_{\mathcal{P}_j \in N_i(t)} d_k^T a_j \ \geq \ d_k^T a_i' \geq \min_{\mathcal{P}_j \in N_i(t)} d_k^T a_j\}.$$

The consensus algorithm in (6.1) corresponds to a specific reasonable utility function. In particular, the set of reasonable actions is the convex hull of the previous actions of his neighbors, i.e.,

$$S_i(a,t) = \{a_i' \in \mathcal{A}_i : a_i' = \sum_{\mathcal{P}_j \in N_i(t)} \omega_{ij} a_j, \ \sum_{\mathcal{P}_j \in N_i(t)} \omega_{ij} = 1, \ \omega_{ij} > 0 \, \forall \, \mathcal{P}_j \in N_i(t)\}.$$

In the present setting, a player's future action need not be in the convex hull of his neighbors' actions.

6.6 Group Based Decision Processes for Potential Games

In this section, we analyze the situation where players are allowed to collaborate with a group of other players when making a decision. In particular we extend the results of SAP to accommodate such a grouping structure.

6.6.1 Spatial Adaptive Play with Group Based Decisions

Consider a potential game with potential function $\phi : \mathcal{A} \to \mathbb{R}$. We will now introduce a variant of SAP to accommodate group based decisions. At each time $t > 0$, a group of players $G \subseteq \mathcal{P}$ is randomly chosen according to a fixed probability distribution $P \in \Delta(2^{\mathcal{P}})$ where $2^{\mathcal{P}}$ denotes the set of subsets of $\mathcal{P}$. We will refer to P_G as the probability that group G will be chosen. We make the following assumption on the group probability distribution.

Assumption 6.6.1 (Completeness). *For any player $\mathcal{P}_i \in \mathcal{P}$ there exists a group $G \subseteq \mathcal{P}$ such that $\mathcal{P}_i \in G$ and $P_G > 0$.*

Once a group is selected, the group is unilaterally allowed to alter it's collective strategy. All players not in the group must repeat their action, i.e., $a_{-G}(t) = a_{-G}(t{-}1)$, where a_G is the action-tuple of all players in the group G, and a_{-G} is the action-tuple of all players not in the group G. The group will be modeled as a single entity with a group utility function $U_G : \mathcal{A} \to \mathbb{R}$ and a collective action set $\mathcal{A}_G = \prod_{\mathcal{P}_i \in G} \mathcal{A}_i$. At time t, the updating group G randomly selects a collective action from $\mathcal{A}_G$ according to the collective strategy $p_G(t) \in \Delta(\mathcal{A}_G)$ where the a_G-th component $p_G^{a_G}(t)$ of the collective strategy is given as

$$p_G^{a_G}(t) = \frac{\exp\{\beta\, U_i(a_G, a_{-G}(t-1))\}}{\sum_{\bar{a}_G \in \mathcal{A}_G} \exp\{\beta\, U_i(\bar{a}_G, a_{-G}(t-1))\}},$$

for some exploration parameter $\beta \geq 0$.

We make the following assumption on the admissible group utility functions:

Assumption 6.6.2 (Group Utility Functions). *Group utility functions must preserve the potential structure of the game, meaning that for any group $G \subseteq \mathcal{P}$, collective group actions $a'_G, a''_G \in \mathcal{A}_G$, and $a_{-G} \in \prod_{\mathcal{P}_i \notin G} \mathcal{A}_i$,*

$$U_G(a''_G, a_{-G}) - U_G(a'_G, a_{-G}) = \phi(a''_G, a_{-G}) - \phi(a'_G, a_{-G}).$$

.

In general, group utility functions need to preserve this condition. However, one can always assign each group a utility that captures the group's marginal contribution to the potential function, i.e., a wonderful life utility as discussed in Section 6.3. This utility assignment guarantees preservation of the potential structure of the game.

We will now show that the convergence properties of the learning algorithm SAP still hold with group based decisions.

Theorem 6.6.1. *Consider a finite n-player potential game with potential function $\phi(\cdot)$, a group probability distribution P satisfying Assumption 6.6.1, and group utility func-*

tions satisfying Assumption 6.6.2. SAP with group based decisions induces a Markov process over the state space $\mathcal{A}$ where the unique stationary distribution $\mu \in \Delta(\mathcal{A})$ is given as

$$\mu(a) = \frac{\exp\{\beta\ \phi(a)\}}{\sum_{\bar{a}\in\mathcal{A}} \exp\{\beta\ \phi(\bar{a})\}},\ \textit{for any } a \in \mathcal{A}. \tag{6.7}$$

Proof. The proof follows along the lines of the proof of Theorem 6.2 in [You98]. By Assumption 6.6.1, the Markov process induced by SAP with group based decisions is irreducible and aperiodic; therefore, the process has a unique stationary distribution. Below, we show that this unique distribution must be (6.7) by verifying that the distribution (6.7) satisfies the detailed balanced equations

$$\mu(a)P_{ab} = \mu(b)P_{ba},$$

for any $a, b \in \mathcal{A}$, where

$$P_{ab} := \mathbf{Pr}\left[a(t) = b | a(t-1) = a\right].$$

Note that there are now several ways to transition from a and b when incorporating group based decisions. Let $\bar{G}(a, b)$ represent the group of players with different actions in a and b, i.e.,

$$\bar{G}(a,b) := \{\mathcal{P}_i \in \mathcal{P} : a_i \neq b_i\}.$$

Let $G(a, b) \subseteq 2^{\mathcal{P}}$ be the complete set of player groups for which the transition from a to b is possible, i.e.,

$$G(a,b) := \{G \in 2^{\mathcal{P}} : \bar{G}(a,b) \subseteq G\}.$$

Since a group $G \in G(a, b)$ has probability P_G of being chosen in any given period, it follows that

$$\mu(a)P_{ab} = \left[\frac{\exp\{\beta\ \phi(a)\}}{\sum_{z\in\mathcal{A}} \exp\{\beta\ \phi(z)\}}\right] \times \left[\sum_{G\in G(a,b)} P_G\ \frac{\exp\{\beta\ U_G(b)\}}{\sum_{\bar{a}_G\in\mathcal{A}_G} \exp\{\beta\ U_G(\bar{a}_G, a_{-G})\}}\right].$$

Letting

$$\lambda_G := \left(\frac{1}{\sum_{z \in \mathcal{A}} \exp\{\beta\ \phi(z)\}}\right) \times \left(\frac{P_G}{\sum_{\bar{a}_G \in \mathcal{A}_G} \exp\{\beta\ U_G(\bar{a}_G, a_{-G})\}}\right),$$

we obtain

$$\mu(a) P_{ab} = \sum_{G \in G(a,b)} \lambda_G \exp\{\beta\phi(a) + \beta U_G(b)\}.$$

Since $U_G(b) - U_G(a) = \phi(b) - \phi(a)$ and $G(a,b) = G(b,a)$, we have

$$\mu(a) P_{ab} = \sum_{G \in G(b,a)} \lambda_G \exp\{\beta\phi(b) + \beta U_G(a)\},$$

which leads us to

$$\mu(a) P_{ab} = \mu(b) P_{ba}.$$

□

6.6.2 Restricted Spatial Adaptive Play with Group Based Decisions

Extending these results to accommodate restricted action sets is straightforward. Let $a(t-1)$ be the action profile at time $t-1$. In this case, the restricted action set for any group $G \subseteq \mathcal{P}$ at time t will be $\mathcal{A}_G(t) = \prod_{\mathcal{P}_i \in G} R_i(a_i(t-1))$. We will state the following theorem without proof to avoid redundancy.

Theorem 6.6.2. *Consider a finite n-player potential game with potential function $\phi(\cdot)$, a group probability distribution P satisfying Assumption 6.6.1, and group utility functions satisfying Assumption 6.6.2. If the restricted action sets satisfy Assumptions 6.3.1 and 6.3.2, then RSAP induces a Markov process over the state space $\mathcal{A}$ where the unique stationary distribution $\mu \in \Delta(\mathcal{A})$ is given as*

$$\mu(a) = \frac{\exp\{\beta\ \phi(a)\}}{\sum_{\bar{a} \in \mathcal{A}} \exp\{\beta\ \phi(\bar{a})\}}, \text{ for any } a \in \mathcal{A}.$$

6.6.3 Constrained Action Sets

The learning algorithms SAP or RSAP with group based decisions induced a Markov process over the entire set $\mathcal{A}$. We will now consider the situation in which each group's action set is constrained, i.e., $\mathcal{A}_G \subset \prod_{\mathcal{P}_i \in G} \mathcal{A}_i$. We will assume that the collective action set of each group is time invariant.

Under this framework, the learning algorithms SAP or RSAP with group based decisions induces a Markov process over the constrained set $\bar{\mathcal{A}} \subseteq \mathcal{A}$ which can be characterized as follows: Let $a(0)$ be the initial actions of all players. If $\bar{a} \in \bar{\mathcal{A}}$ then there exists a sequence of action profiles $a(0) = a^0, a^1, ..., a^n = \bar{a}$ with the condition that for all $k \in \{1, 2, ..., n\}$, $a^k = (a^k_{G_k}, a^{k-1}_{-G_k})$ for a group $G_k \subseteq \mathcal{P}$, where $P_{G_k} > 0$ and $a^k_{G_k} \in \mathcal{A}_{G_k}$. The unique stationary distribution $\mu \in \Delta(\bar{\mathcal{A}})$ is given as

$$\mu(a) = \frac{\exp\{\beta\ \phi(a)\}}{\sum_{\bar{a} \in \bar{\mathcal{A}}} \exp\{\beta\ \phi(\bar{a})\}}, \text{ for any } a \in \bar{\mathcal{A}}. \tag{6.8}$$

6.7 Functional Consensus

In the consensus problem, as described in Section 6.3, the global objective was for all agents to reach consensus. In this section, we will analyze the functional consensus problem where the goal is for all players to reach a specific consensus point which is typically dependent on the initial action of all players, i.e.,

$$\lim_{t \to \infty} a_i(t) = f(a(0)), \ \ \forall \mathcal{P}_i \in \mathcal{P},$$

where $a(0) \in \mathcal{A}$ is the initial action of all players and $f : \mathcal{A} \to \mathbb{R}$ is the desired function. An example of such a function for an n-player consensus problem is

$$f(a(0)) = \frac{1}{n} \sum_{\mathcal{P}_i \in \mathcal{P}} a_i(0),$$

for which the goal would be for all players to agree upon the average of the initial actions of all players. We will refer to this specific functional consensus problem as average consensus.

The consensus algorithm of (6.1) achieves the objective of average consensus under the condition that the interaction graph is connected and the associated weighting matrix, $\Omega = \{\omega_{ij}\}_{\mathcal{P}_i,\mathcal{P}_j\in\mathcal{P}}$, is doubly stochastic. A doubly stochastic matrix is any matrix where all coefficients are nonnegative and all column sums and rows sums are equal to 1. The consensus algorithm takes on the following matrix form

$$a(t+1) = \Omega\, a(t).$$

If Ω is a doubly stochastic matrix, then for any time $t > 0$,

$$\mathbf{1}^T a(t+1) = \mathbf{1}^T \Omega\, a(t) = \mathbf{1}^T a(t).$$

Therefore, the sum of the actions of all players is invariant. Hence, if the players achieve consensus, they must agree upon the average.

In order to achieve any form of functional consensus it is imperative that there exist cooperation amongst the players. Players must agree on how to alter their action each iteration. In the consensus algorithm, this cooperation is induced by the weighting matrix which specifies precisely how a player should change his action each iteration. If a player acted selfishly and unilaterally altered his action, the invariance of the desired function would not be preserved.

6.7.1 Setup: Functional Consensus Problem with Group Based Decisions

Consider the consensus problem with a time invariant undirected interaction graph as described in Section 6.3. To apply the learning algorithm SAP or RSAP with group based decisions to the functional consensus problem one needs to define both the group utility functions and the group selection process.

6.7.2 Group Utility Function

Recall the potential function used for the consensus problem with a time invariant and undirected interaction graph analyzed in Section 6.3,

$$\phi(a) = -(1/2) \sum_{\mathcal{P}_i \in \mathcal{P}} \sum_{\mathcal{P}_j \in N_i} \|a_i - a_j\|.$$

We will assign any group $G \subseteq \mathcal{P}$ the following local group utility function

$$U_G(a) = -(1/2) \sum_{\mathcal{P}_i \in G} \sum_{\mathcal{P}_j \in N_i \cap G} \|a_i - a_j\| - \sum_{\mathcal{P}_i \in G} \sum_{\mathcal{P}_j \in N_i \setminus G} \|a_i - a_j\|. \tag{6.9}$$

An explanation for the $(1/2)$ is to avoid double counting since the interaction graph is undirected. We will now show that this group utility function satisfies Assumption 6.6.2. Before showing this, let N_G denote the neighbors of group G, i.e., $N_G = \bigcup_{\mathcal{P}_i \in G} N_i$. The change in the potential function by switching from $a = (a_G, a_{-G})$ to $a' = (a'_G, a_{-G})$ is

$$\phi(a') - \phi(a) \;=\; -(1/2) \sum_{\mathcal{P}_i \in \mathcal{P}} \sum_{\mathcal{P}_j \in N_i} \big(\|a'_i - a'_j\| - \|a_i - a_j\|\big).$$

For simplicity of notation let $\delta_{ij} = -(1/2)(\|a'_i - a'_j\| - \|a_i - a_j\|)$. The change in the potential can be expressed as

$$\begin{aligned}
\phi(a') - \phi(a) &= \sum_{\mathcal{P}_i \in \mathcal{P}} \sum_{\mathcal{P}_j \in N_i} \delta_{ij}, \\
&= \sum_{\mathcal{P}_i \in N_G} \sum_{\mathcal{P}_j \in N_i} \delta_{ij}, \\
&= \sum_{\mathcal{P}_i \in G} \sum_{\mathcal{P}_j \in N_i \cap G} \delta_{ij} + \sum_{\mathcal{P}_i \in G} \sum_{\mathcal{P}_j \in N_i \setminus G} \delta_{ij} + \sum_{\mathcal{P}_i \in N_G \setminus G} \sum_{\mathcal{P}_j \in N_i} \delta_{ij}, \\
&= \sum_{\mathcal{P}_i \in G} \sum_{\mathcal{P}_j \in N_i \cap G} \delta_{ij} + \sum_{\mathcal{P}_i \in G} \sum_{\mathcal{P}_j \in N_i \setminus G} \delta_{ij} + \sum_{\mathcal{P}_i \in N_G \setminus G} \sum_{\mathcal{P}_j \in N_i \cap G} \delta_{ij}.
\end{aligned}$$

Since the interaction graph is undirected, we know that

$$\sum_{\mathcal{P}_i \in G} \sum_{\mathcal{P}_j \in N_i \setminus G} \delta_{ij} = \sum_{\mathcal{P}_i \in N_G \setminus G} \sum_{\mathcal{P}_j \in N_i \cap G} \delta_{ij},$$

therefore, we can conclude that

$$\begin{aligned}\phi(a') - \phi(a) &= \sum_{\mathcal{P}_i \in G} \big(\sum_{\mathcal{P}_j \in N_i \cap G} \delta_{ij} + 2 \sum_{\mathcal{P}_j \in N_i \setminus G} \delta_{ij} \big) \\ &= U_G(a') - U_G(a).\end{aligned}$$

6.7.3 Group Selection Process and Action Constraints

Let $a(t-1)$ be the action profile at time $t-1$. At time t, one player $\mathcal{P}_i$ is randomly (uniformly) chosen. Rather that updating his action unilaterally, player $\mathcal{P}_i$ first selects a group of players $G \subseteq \mathcal{P}$ which we will assume is the neighbors of player $\mathcal{P}_i$, i.e., $G = N_i$. The group is assigned a group utility function as in (6.9) and a constrained action set $\mathcal{A}_G \subset \prod_{\mathcal{P}_i \in G} \mathcal{A}_i$.

A central question is how can one constrain the group action set, using only location information, such as to preserve the invariance of the desired function f. In this case, we will restrict our attention only to functions where "local" preservation equates to "global" preservation. This means that for each group $G \subseteq \mathcal{P}$ there exists a function $f_G : \mathcal{A}_G \to \mathbb{R}$ such that for any group actions $a'_G, a''_G \in \mathcal{A}_G$

$$f_G(a'_G) = f_G(a''_G) \ \Rightarrow \ f(a'_G, a_{-G}) = f(a''_G, a_{-G}), \ \forall a_{-G} \in \prod_{\mathcal{P}_i \notin G} \mathcal{A}_i.$$

Examples of functions that satisfy this constraint are

$$\begin{aligned} f_G(a) = \frac{1}{|G|} \sum_{\mathcal{P}_i \in G} a_i \ &\Rightarrow \ f(a) = \frac{1}{|\mathcal{P}|} \sum_{\mathcal{P}_i \in \mathcal{P}} a_i, \\ f_G(a) = \max_{\mathcal{P}_i \in G} a_i \ &\Rightarrow \ f(a) = \max_{\mathcal{P}_i \in \mathcal{P}} a_i, \\ f_G(a) = \min_{\mathcal{P}_i \in G} a_i \ &\Rightarrow \ f(a) = \min_{\mathcal{P}_i \in \mathcal{P}} a_i. \end{aligned}$$

In each of these examples, the structural form of f and f_G is equivalent. There may exist alternative functions where this is not required.

6.7.4 Illustration

We will illustrate this approach by solving the average consensus problem on the example developed in Section 6.3.4. Given the initial configuration, all players should agree upon the action $(5, 5)$. We will solve this average consensus problem using the learning algorithm binary RSAP with group based decisions. However, we will omit the non-convex obstruction in this illustration. This omission is not necessary, but rather convenient for not having to verify the properties of the constrained action set, i.e., is consensus even possible, and Assumption 6.3.2 for the group action sets.

Figure 6.3 illustrates the evolution of each player's actions using the stochastic learning algorithm binary RSAP with group based decisions and an increasing β coefficient, $\beta(t) = 1.5 + t(2/1000)$.

6.8 Illustrative Examples

In this section we will develop two examples to further illustrate the wide range applicability of the theory developed in this chapter. The first problem we will consider is the dynamic sensor allocation problem. Lastly, we will demonstrate how this theory can be used to solve a popular mathematical puzzle called Sudoku.

6.8.1 Dynamic Sensor Coverage Problem

We consider the dynamic sensor coverage problem described in [LC05c] and references therein. The goal of the sensor coverage problem is to allocate a fixed number of sensors across a given "mission space" to maximize the probability of detecting a particular event.

We will divide the mission space into a finite set of sectors denoted as S. There

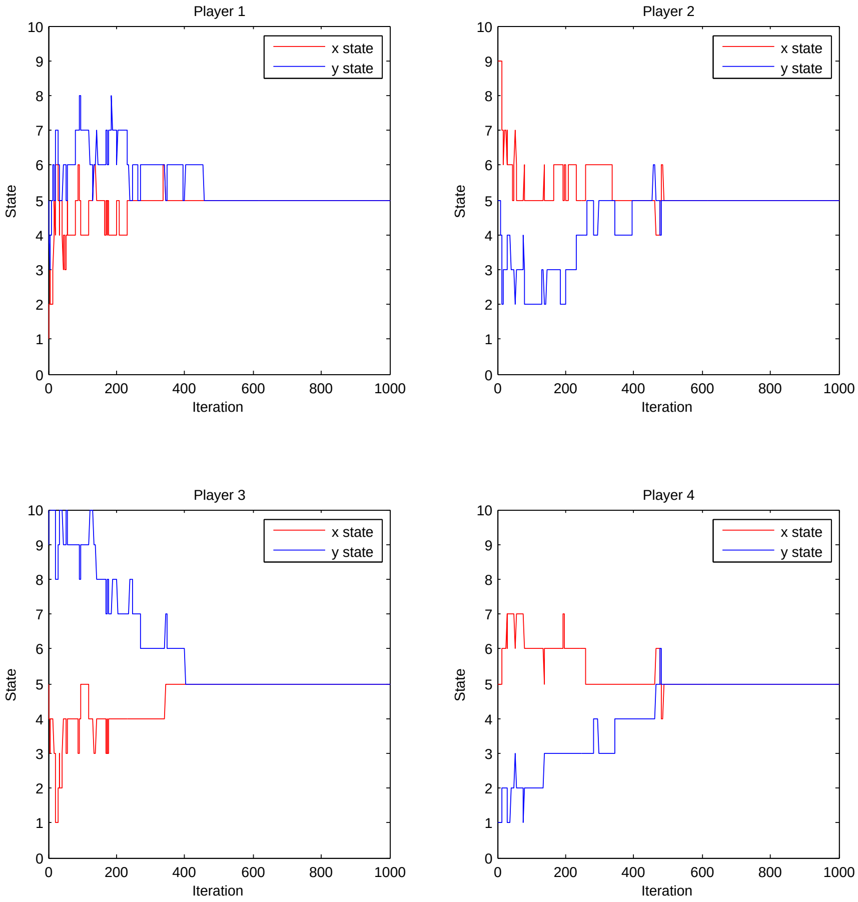


Figure 6.3: Evolution of Each Player's Action in the Average Consensus Problem

exists an events density function, or relative reward function, $R(s)$, defined over S. We will assume that $R(s) \geq 0, \forall s \in S$ and $\sum_{s \in S} R(s) = 1$. In the application of enemy submarine tracking, $R(s)$ could be defined as the a priori probability that an enemy submarine is situated in sector s. The mission space and associated reward function that we will use in this section is illustrated in Figure 6.4.

There are a finite number of autonomous sensors denoted as $\mathcal{P} = \{\mathcal{P}_1, ..., \mathcal{P}_n\}$ allocated to the mission space. Each sensor $\mathcal{P}_i$ can position itself in any particular

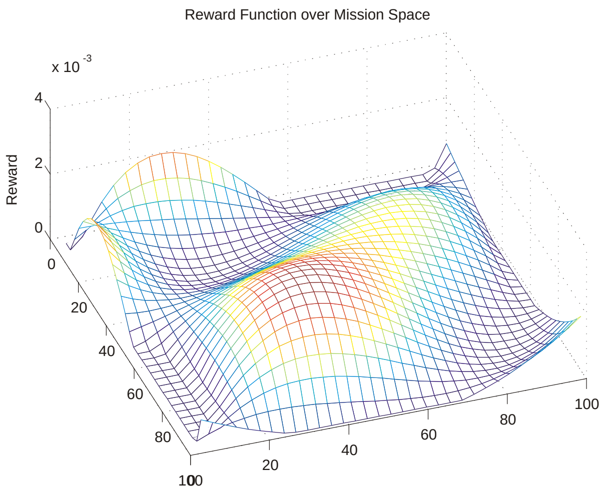


Figure 6.4: Illustration of Reward Function Over Mission Space

sector, i.e., the action set of sensor $\mathcal{P}_i$ is $\mathcal{A}_i = S$. Furthermore, each sensor has limited sensing and moving capabilities. If an event occurs in sector s, the probability of sensor $\mathcal{P}_i$ detecting the event given his current location a_i is denoted as $p_i(s, a_i)$. Typically, each sensor has a finite sensing radius, r_i, where the probability of detection obeys the following:

$$\|s - a_i\| < r_i \Leftrightarrow p_i(s, a_i) > 0.$$

An example of the sensing and moving capabilities of a particular sensor is illustrated in Figure 6.5.

For a given joint action profile $a = \{a_1, ..., a_n\}$, the joint probability of detecting an event in sector s is given by

$$P(s, a) = 1 - \prod_{\mathcal{P}_i \in \mathcal{P}} [1 - p_i(s, a_i)].$$

In general a global planner would like the sensors to allocate themselves in such a fashion as to maximize the following potential function

$$\phi(a) = \sum_{s \in S} R(s) P(s, a).$$

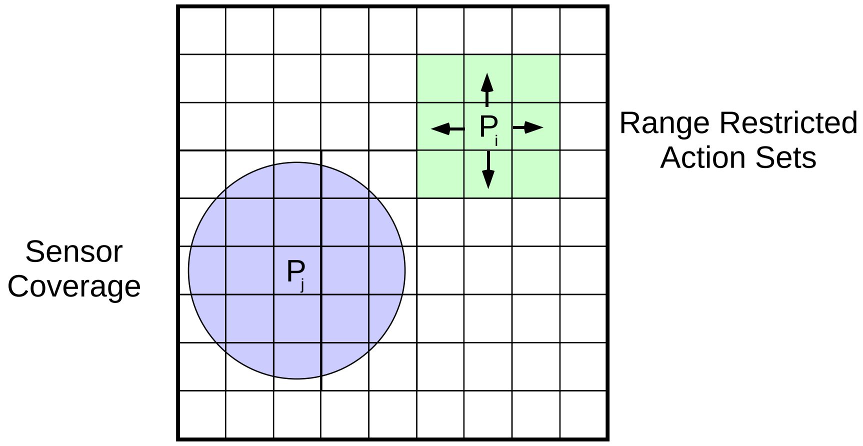


Figure 6.5: Illustration of Sensor Coverage and Range Restricted Action Sets of a Particular Sensor

One way to accomplish such an objective is to assign each sensor a utility function that is appropriately aligned with the global objective function as was the case in the consensus problem. One option is to just assign each sensor the global objective, i.e.,

$$U_i(a) = \phi(a).$$

Under this scenario, we have a potential game and one could use a learning algorithm like SAP or RSAP to guarantee that the sensors allocate themselves in a configuration that maximizes the global objective. However, this particular choice of utility functions require each sensor to be knowledgable of the locations and capabilities of all other sensors. To avoid this requirement, we will assign each sensor a Wonderful Life Utility [AMS07, WT99]. The utility of sensor $\mathcal{P}_i$ given any action profile $a \in \mathcal{A}$ is now

$$U_i(a) = \phi(a_i, a_{-i}) - \phi(a_i^0, a_{-i}), \tag{6.10}$$

where the action a_i^0 is defined as the *null* action, which is equivalent to sensor $\mathcal{P}_i$ turning off all sensing capabilities. The term $\phi(a_i^0, a_{-i})$ captures the value of the allocation of all sensors other than sensor $\mathcal{P}_i$. Therefore, the utility of sensor $\mathcal{P}_i$ for an action profile a is defined as his *marginal* contribution to the global objective. This means that a sensor now can evaluate his utility using only local information. Furthermore, the Wonderful Life Utility assignment preserves the potential game structure [AMS07, WT99], meaning that SAP or RSAP can now be implemented with the sensors using only local information to guarantee that the sensors allocate themselves in a desirable configuration.

In the following simulation we have the mission space and reward function as illustrated in Figure 6.4. The mission space is $S = \{1, 2, ..., 100\} \times \{1, 2, ..., 100\}$ and the reward function satisfies $\sum_{s \in S} R(s) = 1$. We have 18 different autonomous sensors, 6 with a sensing radius of 6, 6 with a sensing radius of 12, and 6 with a sensing radius of 18. For simplicity, each sensor will have prefect sensing capabilities within its sensing radius, namely for any sector s satisfying $\|s - a_i\| < r_i$, then $p_i(s, a_i) = 1$. Each sensor is endowed with the WLU as expressed in (6.10). All 18 sensors originally started at the location $(1, 1)$ and each sensor has range restricted action sets as illustrated in Figure 6.5. We ran the binary RSAP with $\beta = 0.6$. Figure 6.6 illustrates a snapshot of the sensors configuration at the final iteration. Figure 6.7 illustrates the evolution of the potential function over the mission.

6.8.2 Sudoku

Our final illustration of the broad applicability of potential games is the well known mathematical puzzle of Sudoku. An example of a Sudoku puzzle is shown in Figure 6.8. The objective is to fill a 9x9 grid so that each column, each row, and each of the nine 3x3 boxes contains the digits from 1 to 9. The puzzle setter provides a partially completed grid (blue boxes) which cannot be changed.

We will now illustrate that Sudoku is exactly a potential game when the players, action sets, and utility functions are designed appropriately. We will view each of the

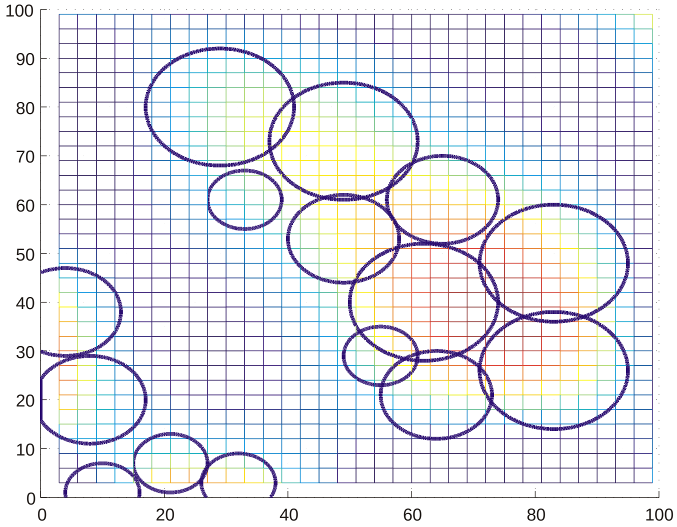


Figure 6.6: Final Allocation of Sensors over Mission Space

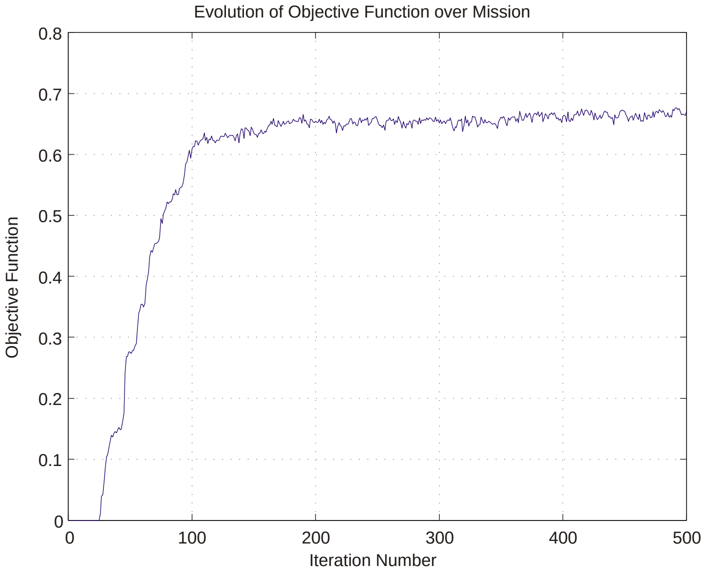
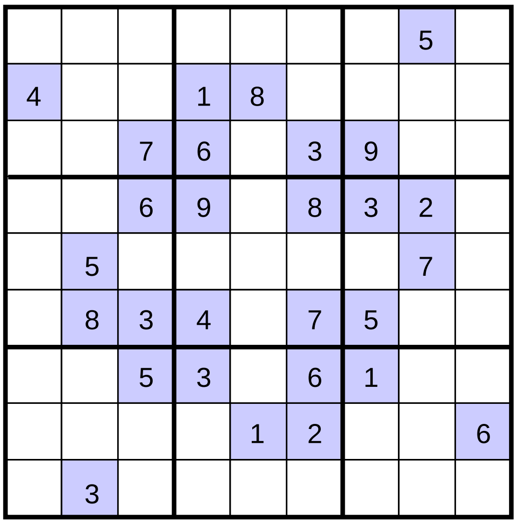


Figure 6.7: Evolution of Potential Function over Mission



							5	
4			1	8				
		7	6		3	9		
		6	9		8	3	2	
	5						7	
	8	3	4		7	5		
		5	3		6	1		
				1	2			6
	3							

Figure 6.8: Illustration of a Sudoku Puzzle

empty boxes as a self interested player $\mathcal{P}_i$ with action set $\mathcal{A}_i = \{1, 2, ..., 9\}$. Each player will be assigned the utility function

$$U_i(a) \ := \ \sum_{\mathcal{P}_j \in N_i^R} I\{a_i = a_j\} + \sum_{\mathcal{P}_j \in N_i^C} I\{a_i = a_j\} + \sum_{\mathcal{P}_j \in N_i^B} I\{a_i = a_j\},$$

where N_i^R, N_i^C, and N_i^B are the row, column and box neighbors of player $\mathcal{P}_i$ and $I\{\cdot\}$ is the usual indication function. An illustration of the neighbor sets of player $\mathcal{P}_1$ is highlighted in Figure 6.9, where the the green boxes indicate the row neighbors, red boxes indicate the column neighbors, and yellow boxes indicate the box neighbors. Note that in this framework, unlike with the consensus problem, each player $\mathcal{P}_i$ is not a neighbor of himself.

To simplify the notation, we define the following function: for each player $\mathcal{P}_i$ and for any player set $\bar{\mathcal{P}} \subseteq \mathcal{P}$, let

$$n_i(a, \bar{\mathcal{P}}) := \sum_{\mathcal{P}_j \in \bar{\mathcal{P}}} I\{a_i = a_j\}.$$

This function computes the number of players with the same action as player $\mathcal{P}_i$ in the

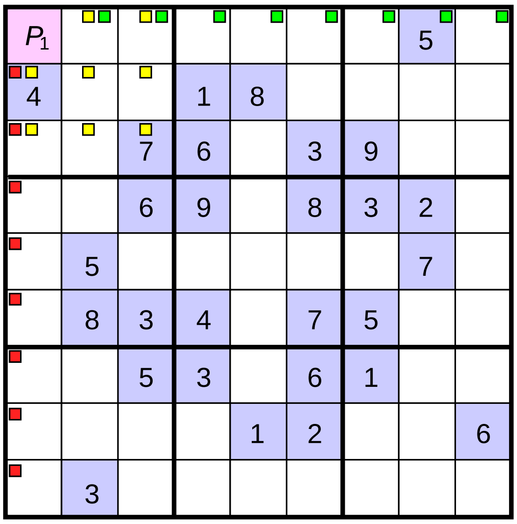


Figure 6.9: Illustration of Neighbor Sets for a Player's Utility Function in a Sudoku Puzzle

set $\bar{\mathcal{P}}$. Using this function, we will express the utility of player $\mathcal{P}_i$ as

$$U_i(a) = n_i(a, N_i^R) + n_i(a, N_i^C) + n_i(a, N_i^B).$$

We will now show that the Sudoku game with utilities defined as above is a potential game with potential function

$$\phi(a) \quad = \quad 1/2 \sum_{\mathcal{P}_i \in \mathcal{P}} U_i(a).$$

To simplify the analysis, we will break up the potential function as

$$\phi(a) = \phi^R(a) + \phi^C(a) + \phi^B(a),$$

where

$$\begin{aligned} \phi^R(a) &= 1/2 \sum_{\mathcal{P}_i \in \mathcal{P}} n_i(a, N_i^R), \\ \phi^C(a) &= 1/2 \sum_{\mathcal{P}_i \in \mathcal{P}} n_i(a, N_i^C), \\ \phi^B(a) &= 1/2 \sum_{\mathcal{P}_i \in \mathcal{P}} n_i(a, N_i^B). \end{aligned}$$

Let $a', a'' \in Y$ be any two action profiles that differ by a unilateral deviation, i.e., $a'_i \neq a''_i$ and $a'_{-i} = a''_{-i}$ for some player $\mathcal{P}_i \in \mathcal{P}$. The change in $\phi^R(\cdot)$ is

$$\begin{aligned}
2(\phi^R(a') - \phi^R(a'')) &= \sum_{\mathcal{P}_i \in \mathcal{P}} n_i(a', N_i^R) - n_i(a'', N_i^R), \\
&= n_i(a', N_i^R) - n_i(a'', N_i^R) + \sum_{\mathcal{P}_j \in N_i^R} n_j(a', N_j^R) - n_j(a'', N_j^R), \\
&= n_i(a', N_i^R) - n_i(a'', N_i^R) + \sum_{\mathcal{P}_j \in N_i^R} n_j(a', \mathcal{P}_i) - n_j(a'', \mathcal{P}_i), \\
&= n_i(a', N_i^R) - n_i(a'', N_i^R) + \sum_{\mathcal{P}_j \in N_i^R} n_i(a', \mathcal{P}_j) - n_i(a'', \mathcal{P}_j), \\
&= n_i(a', N_i^R) - n_i(a'', N_i^R) + n_i(a', N_i^R) - n_i(a'', N_i^R), \\
&= 2(n_i(a', N_i^R) - n_i(a'', N_i^R)).
\end{aligned}$$

One could repeat this analysis for $\phi^C(\cdot)$ and $\phi^B(\cdot)$ to show that

$$\phi(a') - \phi(a'') = U_i(a') - U_i(a'').$$

Therefore the Sudoku game is in fact a potential game. Furthermore, the potential function is always nonnegative, and achieves the value of 0 if and only if the Sudoku puzzle has been solved. Therefore, all solutions to the Sudoku puzzles are in fact Nash equilibria of the Sudoku game. However, much like the consensus problem, there may exist suboptimal Nash equilibria.

To solve the Sudoku puzzle we will use the learning algorithm SAP as described in Section 6.3.2. We let the β coefficient increase as $\beta(t) = t/5000$. Figure 6.10 shows the evolution of the potential function during the SAP learning process. One can see that the potential function achieves the value of 0 after approximately 17,000 iterations which means that the puzzle has been solved. To verify, the final joint action profile is illustrated in Figure 6.11.

To further illustrate the applicability of SAP, we simulated SAP on a Sudoku puzzle

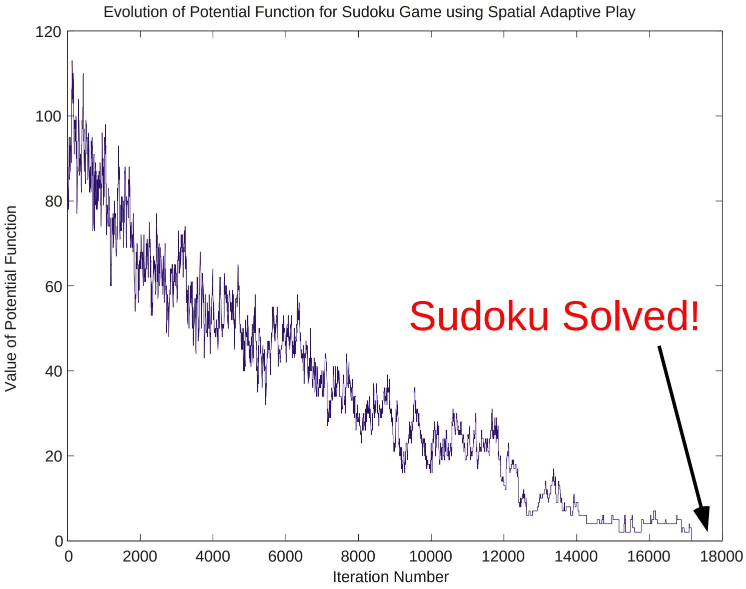
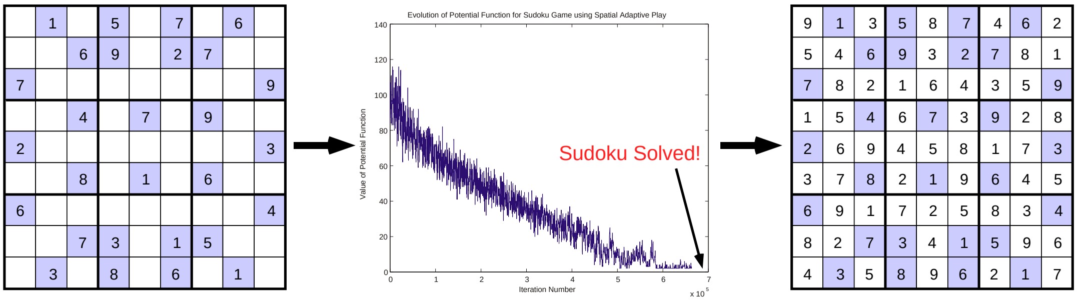


Figure 6.10: Evolution of Potential Function in Sudoku Puzzle Under the Learning Algorithm Spatial Adaptive Play

3	6	8	7	2	9	4	5	1
4	9	2	1	8	5	7	6	3
5	1	7	6	4	3	9	8	2
1	7	6	9	5	8	3	2	4
9	5	4	2	3	1	6	7	8
2	8	3	4	6	7	5	1	9
8	2	5	3	9	6	1	4	7
7	4	9	5	1	2	8	3	6
6	3	1	8	7	4	2	9	5

Figure 6.11: The Completed Sudoku Puzzle

classified as *very hard*. Once again, a solution to the puzzle was found as illustrated in Figure 6.12.



	1		5		7		6	
		6	9		2	7		
7								9
		4		7		9		
2								3
		8		1		6		
6								4
		7	3		1	5		
	3		8		6		1	

9	1	3	5	8	7	4	6	2
5	4	6	9	3	2	7	8	1
7	8	2	1	6	4	3	5	9
1	5	4	6	7	3	9	2	8
2	6	9	4	5	8	1	7	3
3	7	8	2	1	9	6	4	5
6	9	1	7	2	5	8	3	4
8	2	7	3	4	1	5	9	6
4	3	5	8	9	6	2	1	7

Figure 6.12: Spatial Adaptive Play on a Sudoku Puzzle Classified as Very Hard

It is important to note that while it took many iterations to solve the Sudoku puzzles, the algorithm of SAP was applied in its original form. We firmly believe that the algorithm could be modified to decrease computation time. For example, a player's action set could be reduced with knowledge of the board. In particular, the action set of player $\mathcal{P}_1$ in Figure 6.9 could initially have been set as $\mathcal{A}_1 = \{1, 2, 3, 6, 7, 8, 9\}$.

6.9 Concluding Remarks

We have proposed a game theoretic approach to cooperative control by highlighting a connection between cooperative control problems and potential games. We introduced a new class of games and enhanced existing learning algorithms to broaden the applicability of game theoretic methods in cooperative control setting. We demonstrated that one could successfully implement game theoretic methods on the cooperative control problem of consensus in a variety of settings. While the main example used was the consensus problem, the results in Theorems 6.3.1, 6.4.1, and 6.6.1 and the notion of a sometimes weakly acyclic game is applicable to a broader class of games as well as other cooperative control problems.

CHAPTER 7

Conclusions

This dissertation focused on dealing with the distributed nature of decision making and information processing through a non-cooperative game-theoretic formulation. The emphasis was on simple learning algorithms that guarantee convergence to a Nash equilibrium.

We analyzed the long-term behavior of a large number of players in large-scale games where players are limited in both their observational and computational capabilities. In particular, we analyzed a version of JSFP and showed that it accommodates inherent player limitations in information gathering and processing. Furthermore, we showed that JSFP has guaranteed convergence to a pure Nash equilibrium in all generalized ordinal potential games, which includes but is not limited to all congestion games, when players use some inertia either with or without exponential discounting of the historical data. Furthermore, we introduced a modification of the traditional no-regret algorithms that (i) exponentially discounts the memory and (ii) brings in a notion of inertia in players' decision process. We showed how these modifications can lead to an entire class of regret based algorithms that provide convergence to a pure Nash equilibrium in any weakly acyclic game.

The method of proof used for JSFP and the regret based dynamics relies on inertia to derive a positive probability of a single player seeking to make an utility improvement, thereby increasing the potential function. This suggests a convergence rate that is exponential in the game size, i.e., number of players and actions. It should be noted that inertia is simply a proof device that assures convergence for generic potential games. The proof provides just one out of multiple paths to convergence. The simulations reflect that convergence can be much faster. Indeed, simulations suggest that convergence is possible even in the absence of inertia. Furthermore, recent work [HM06] suggests that convergence rates of a broad class of distributed learning processes can be exponential in the game size as well, and so this seems to be a limitation in the framework of distributed learning rather than any specific learning process (as opposed to centralized algorithms for computing an equilibrium).

We also analyzed the long-term behavior of a large number of players in large-scale games where players only have access to the action they played and the utility they received. Our motivation for this information restriction is that in many engineered systems, the functional forms of utility functions are not available, and so players must adjust their strategies through an adaptive process using only payoff measurements. In the dynamic processes defined here, there is no explicit cooperation or communication between players. One the one hand, this lack of explicit coordination offers an element of robustness to a variety of uncertainties in the strategy adjustment processes. Nonetheless, an interesting future direction would be to investigate to what degree explicit coordination through limited communications could be beneficial.

In this payoff based setting, players are no longer capable of analyzing the utility they would have received for alternative action choices as required in the regret based algorithms and JSFP. We introduced Safe Experimentation dynamics for identical interest games, Simple Experimentation dynamics for weakly acyclic games with noise-free utility measurements, and Sample Experimentation dynamics for weakly acyclic games with noisy utility measurements. For all three settings, we have shown that for sufficiently large times, the joint action taken by players will constitute a Nash equilibrium. Furthermore, we have shown how to guarantee that a collective objective in a congestion game is a (non-unique) Nash equilibrium.

Lastly, we proposed a game theoretic approach to cooperative control by highlighting a connection between cooperative control problems and potential games. We introduced a new class of games and enhanced existing learning algorithms to broaden the applicability of game theoretic methods in the cooperative control setting. We demonstrated that one could successfully implement game theoretic methods on several cooperative control problems including consensus, dynamic sensor allocation, and distributing routing over a network. Furthermore, we even demonstrated how the mathematical puzzle of Sudoku can be modeled as a potential game and solved in a distributed fashion using the learning algorithms discussed in this dissertation.

In summary, this dissertation illustrated a connection between the fields of learning in games and cooperative control and developed several suitable learning algorithms for a wide variety of cooperative control problems. There remains several interesting and challenging directions for future research.

Equilibrium Selection and Utility Design:

One problem regarding a game theoretic formulation of a multi-agent system is the existence of multiple Nash equilibria, not all of which are desirable operating conditions. Is it possible to develop a methodology for designing agent utilities/objectives and to derive implementable learning algorithms that guarantee the agents' collective behavior converges to a desirable Nash equilibrium? For example, the potential game formulation of the consensus problem had suboptimal Nash equilibria, i.e., Nash equilibria that did not represent consensus points. The existence of these suboptimal Nash equilibria required the use of a stochastic learning algorithm such as SAP or RSAP to guarantee reaching a desirable Nash equilibrium. However, when we modeled the consensus problem as a sometimes weakly acyclic game and properly designed the utilities we were able to effectively eliminate these suboptimal Nash equilibria. Can this be accomplished for more general cooperative control problems?

Learning Algorithms for Stochastic Games:

In many cooperative control problems players are inherently faced with a notion of state dependent action sets and objectives. Stochastic games, which generalize Markov decision processes to multiple decision makers, emerge as the most natural framework to study such cooperative systems. An important research direction is understand to applicability of Markov games for cooperative control problems and to develop simple computational learning algorithms for stochastic games with guaranteed convergence results. We believe that the notion of sometimes weakly acyclic game is an initial step in the direction or Markov games.

Learning Algorithms with Time Guarantees:

One open issue with regarding the applicability of the learning algorithms discussed in this paper is time complexity. Roughly speaking, how long will it take the agents to reach some form of a desirable operating condition? One question that has relevance is whether non-stochastic learning algorithms, such as JSFP and regret based algorithms, have computational advantage over stochastic learning algorithms, such as SAP or RSAP. If the answer to this question is an affirmative, than the notion of utility design plays an even more important role in the applicability of these learning algorithms for controlling multi-agent systems.

REFERENCES

[AMS07] G. Arslan, J. R. Marden, and J. S. Shamma. "Autonomous Vehicle-Target Assignment: A Game Theoretical Formulation." *ASME Journal of Dynamic Systems, Measurement and Control*, 2007. to appear.

[AS04] G. Arslan and J. S. Shamma. "Distributed convergence to Nash equilibria with local utility measurements." In *43rd IEEE Conference on Decision and Control*, pp. 1538–1543, 2004.

[BEL06] A. Blum, E. Evan-Dar, and K. Ligett. "On Convergence to Nash Equilibria of Regret-Minimizing Algorithms in Routing Games." In *Symposium on Principles of Distributed Computing (PODC)*, 2006.

[BHO05] V. D. Blondel, J. M. Hendrickx, A. Olshevsky, and J. N. Tsitsiklis. "Convergence in multiagent coordination, consensus, and flocking." In *IEEE Conference on Decision and Control*, 2005.

[BK03] V. S. Borkar and P. R. Kumar. "Dynamic Cesaro-Wardrop equilibration in networks." *IEEE Transactions on Automatic Control*, **48**(3):382–396, 2003.

[BL85] M. Ben-Akiva and S. Lerman. *Discrete-Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA, 1985.

[Bow04] M. Bowling. "Convergence and No-Regret in Multiagent Learning." In *Neural Information Processing Systems Conference (NIPS)*, 2004.

[BP05] B. Banerjee and J. Peng. "Efficient No-regret Multiagent Learning." In *The 20th National Conference on Artificial Intelligence (AAAI-05)*, 2005.

[BPK91] M. Ben-Akiva, A. de Palma, and I. Kaysi. "Dynamic network models and driver information systems." *Transportation Research A*, **25A**:251–266, 1991.

[Bra68] D. Braess. "Uber ein Paradoxen der Verkehrsplanning." *Unternehmensforschung*, **12**:258–268, 1968.

[BT96] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

[FK93] D. Fudenberg and D. Kreps. "Learning mixed equilibria." *Games and Economic Behavior*, **5**:320–367, 1993.

[FL98] D. Fudenberg and D. K. Levine. *The Theory of Learning in Games*. MIT Press, Cambridge, MA, 1998.

[FRV06] S. Fischer, H. Raecke, and B. Voecking. "Fast convergence to Wardrop equilibria by adaptive sampling methods." In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pp. 653–662, 2006.

[FT91] D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, Cambridge, MA, 1991.

[FV04] S. Fischer and B. Vocking. "The evolution of selfish routing." In *Proceedings of the 12th European Symposium on Algorithms (ESA '04)*, pp. 323–334, 2004.

[FV05] S. Fischer and B. Voecking. "Adaptive routing with stale information." In *Proceedings of the 24th Annual ACM Symposium on Principles of Distributed Computing*, pp. 276–283, 2005.

[FY06] D. P. Foster and H. P. Young. "Regret testing: Learning to play Nash equilibrium without knowing you have an opponent." *Theoretical Economics*, **1**:341–367, 2006.

[Ger94] S. B. Gershwin. *Manufacturing Systems Engineering*. Prentice-Hall, 1994.

[GJ03] A. Greenwald and A. Jafari. "A General Class of No-Regret Learning Algorithms and Game-Theoretic Equilibria." In *Conference on Learning Theory (COLT)*, pp. 2–12, 2003.

[GL] F. Germano and G. Lugosi. "Global convergence of Foster and Young's regret testing." *Games and Economic Behavior*. forthcoming.

[Gor05] G. J. Gordon. "No-regret algorithms for structured prediction problems." Technical Report CMU-CALD-05-112, Department of Machine Learning at Carnegie Mellon, 2005.

[GSM05] A. Ganguli, S. Susca, S. Martinez, F. Bullo, and J. Cortes. "On collective motion in sensor networks: sample problems and distributed algorithms." In *Proceedings of the 44th IEEE Conference on Decision and Control*, pp. 4239–4244, Seville, Spain, December 2005.

[Har05] S. Hart. "Adaptive Heuristics." *Econometrica*, **73**(5):1401–1430, 2005.

[HM00] S. Hart and A. Mas-Colell. "A simple adaptive procedure leading to correlated equilibrium." *Econometrica*, **68**:1127–1150, 2000.

[HM01] S. Hart and A. Mas-Colell. "A general class of adaptative strategies." *Journal of Economic Theory*, **98**:26–54, 2001.

[HM03a] S. Hart and A. Mas-Colell. "Regret based continuous-time dynamics." *Games and Economic Behavior*, **45**:375–394, 2003.

[HM03b] S. Hart and A. Mas-Colell. "Uncoupled dynamics do not lead to Nash equilibrium." *American Economic Review*, **93**(5):1830–1836, 2003.

[HM06] S. Hart and Y. Mansour. "The communication complexity of uncoupled nash equilibrium procedures." Technical Report DP-419, The Hebrew University of Jerusalem, Center for Rationality, April 2006.

[HS98] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, UK, 1998.

[HS04] S. Huck and R. Sarin. "Players with limited memory." *Contributions to Theoretical Economics*, **4**(1), 2004.

[JGD01] A. Jafari, A. Greenwald, D., and G. Ercal. "On No-Regret Learning, Fictitious Play, and Nash Equilibrium." In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pp. 226–233, 2001.

[JLM03] A. Jadbabaie, J. Lin, and A. S. Morse. "Coordination of groups of mobile autonomous agents using nearest neighbor rules." *IEEE Transaction on Automatic Control*, **48**(6):988–1001, June 2003.

[KBS06] A. Kashyap, T. Basar, and R. Srikant. "Consensus with Quantized Information Updates." In *45th IEEE Conference on Decision and Control*, pp. 2728–2733, 2006.

[KV05] A. Kalai and S. Vempala. "Efficient algorithms for online decision problems." *Journal of Computer and System Sciences*, **71**(3):291–307, 2005.

[LC03] D. Leslie and E. Collins. "Convergent multiple-timescales reinforcement learning algorithms in normal form games." *Annals of Applied Probability*, **13**:1231–1251, 2003.

[LC05a] D. Leslie and E. Collins. "Generalised weakened fictitious play." *Games and Economic Behavior*, **56**:285–298, 2005.

[LC05b] D. Leslie and E. Collins. "Individual Q-learning in normal form games." *SIAM Journal on Control and Optimization*, **44**(2), 2005.

[LC05c] W. Li and C. G. Cassandras. "Sensor Networks and Cooperative Control." *European Journal of Control*, 2005. to appear.

[LES05] T. Lambert, M. Epelman, and R. Smith. “A Fictitious Play Approach to Large-Scale Optimization.” *Operations Research*, **53**(3):477–489, 2005.

[MAS05] J. R. Marden, G. Arslan, and J. S. Shamma. “Joint Strategy Fictitious Play with Inertia for Potential Games.” In *Proceedings of the 44th IEEE Conference on Decision and Control*, pp. 6692–6697, December 2005. Submitted to *IEEE Transactions on Automatic Control*.

[MAS07a] J. R. Marden, G. Arslan, and J. S. Shamma. “Connections Between Cooperative Control and Potential Games Illustrated on the Consensus Problem.” In *Proceedings of the 2007 European Control Conference (ECC ’07)*, July 2007. to appear.

[MAS07b] J. R. Marden, G. Arslan, and J. S. Shamma. “Regret Based Dynamics: Convergence in Weakly Acyclic Games.” In *Proceedings of the 2007 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Honolulu, Hawaii, May 2007.

[Mil04] I. Milchtaich. “Social optimality and cooperation in nonatomic congestion games.” *Journal of Economic Theory*, **114**(1):56–87, 2004.

[Mor04] L. Moreau. “Stability of Continuous-Time Distributed Consensus Algorithms.” In *43rd IEEE Conference on Decision and Control*, pp. 3998–4003, 2004.

[MS96a] D. Monderer and L. S. Shapley. “Fictitious play property for games with identical interests.” *Journal of Economic Theory*, **68**:258–265, 1996.

[MS96b] D. Monderer and L. S. Shapley. “Potential Games.” *Games and Economic Behavior*, **14**:124–143, 1996.

[MS97] D. Monderer and A. Sela. “Fictitious play and no-cycling conditions.” Technical report, 1997.

[MS07] S. Mannor and J.S. Shamma. “Multi-agent Learning for Engineers.” 2007. forthcoming special issue in *Artificial Intelligence*.

[MYA07] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma. “Payoff Based Dynamics for Multi-Player Weakly Acyclic Games.” *SIAM Journal of Control and Optimization*, 2007. submitted to.

[OFM07] R. Olfati-Saber, J. A. Fax, and R. M. Murray. “Consensus and Cooperation in Networked Multi-Agent Systems.” In *Proceedings of the IEEE*, January 2007. to appear.

[OM03] R. Olfati-Saber and R. M. Murray. “Consensus Problems in Networks of Agents with Switching Topology and Time-Delays.” *IEEE Transaction on Automatic Control*, **49**(6), June 2003.

[Ros73] R. W. Rosenthal. “A Class of Games Possessing Pure-Strategy Nash Equilibria.” *Int. J. Game Theory*, **2**:65–67, 1973.

[Rou03] Tim Roughgarden. “The price of anarchy is independent of the network topology.” *Journal of Computer and System Sciences*, **67**(2):341–364, 2003.

[SA05] J. S. Shamma and G. Arslan. “Dynamic fictitious play, dynamic gradient play, and distributed convergence to Nash equilibria.” *IEEE Transactions on Automatic Control*, **50**(3):312–327, 2005.

[Sam97] L. Samuelson. *Evolutionary Games and Equilibrium Selection*. MIT Press, Cambridge, MA, 1997.

[San02] W. Sandholm. “Evolutionary Implementation and Congestion Pricing.” *Review of Economic Studies*, **69**(3):667–689, 2002.

[SB98] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, MA, 1998.

[SPG07] Y. Shoham, R. Powers, and T. Grenager. “If multi-agent learning is the answer, what is the question?” forthcoming special issue in *Artificial Intelligence*, 2007.

[TBA86] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans. “Distributed Asynchronous Deterministic and Stochastic Gradient Optimization Algorithms.” *IEEE Transactions on Automatic Control*, **35**(9):803–812, 1986.

[War52] J. G. Wardrop. “Some theoretical aspects of road traffic research.” In *Proceedings of the Institute of Civil Engineers*, volume I, pt. II, pp. 325–378, London, Dec. 1952.

[Wei95] J.W. Weibull. *Evolutionary Game Theory*. MIT Press, Cambridge, MA, 1995.

[WT99] D. Wolpert and K. Tumor. “An overview of collective intelligence.” In J. M. Bradshaw, editor, *Handbook of Agent Technology*. AAAI Press/MIT Press, 1999.

[XB04] L. Xiao and S. Boyd. “Fast linear iterations for distributed averaging.” *Systems and Control Letters*, 2004.

[XB05] L. Xiao and S. Boyd. “A scheme for robust distributed sensor fusion based on average consensus.” In *Information processing in sensor networks*, 2005.

[You93] H. P. Young. “The Evolution of Conventions.” *Econometrica*, **61**(1):57–84, January 1993.

[You98] H. P. Young. *Individual Strategy and Social Structure*. Princeton University Press, Princeton, NJ, 1998.

[You05] H. P. Young. *Strategic Learning and its Limits*. Oxford University Press, 2005.